

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Axolotl genome

Salamander sequence sheds light on development and regeneration **PAGES 34 & 50**

ECOLOGY

QUICK-CHANGE ARTISTS

Rapid evolution reshapes ecological research

PAGE 19

ORGANIC CHEMISTRY

CARBYNE HARVESTING

Reactive carbon species made available for synthesis

PAGES 36 & 86

VIROLOGY

DENIZENS OF THE DEEP

Elusive family of tail-free marine viruses identified

PAGES 38 & 118

NATURE.COM/NATURE

1 February 2018 £10

Vol. 554, No. 7690

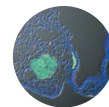


THIS WEEK

EDITORIALS

PUBLISHING Nature journals tighten the rules on conflicts of interest **p.6**

WORLD VIEW Assess the social dynamics of peer-review panels **p.7**



CANCER Nanoparticles made of sugar target tumour cells **p.9**

Lucky science

Scientists often herald the role of serendipity in research. A project in Britain aims to test the popular idea with evidence.

Science-fiction writer Isaac Asimov is widely credited with saying that “the most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ but ‘That’s funny.’” Scientific folklore is full of tales of accidental discovery, from the stray Petri dish that led Alexander Fleming to discover penicillin to Wilhelm Röntgen’s chance detection of X-rays while tinkering with a cathode-ray tube.

That knowledge often advances through serendipity is how scientists, sometimes loudly, justify the billions of dollars that taxpayers plough into curiosity-driven research each year. And it is the reason some argue that increasing government efforts to control research — with an eye to driving greater economic or social impact — are at best futile and at worst counterproductive.

But just how important is serendipity to science? Scientists debating with policymakers have long relied on anecdotal evidence. Studies rarely try to quantify how much scientific progress was truly serendipitous, how much that cost or the circumstances in which it emerged.

Serendipity can take on many forms, and its unwieldy web of cause and effect is difficult to constrain. Data are not available to track it in any meaningful way. Instead, academic research has focused on serendipity in science as a philosophical concept.

The European Research Council aims to change that. It has given biochemist-turned-social-scientist Ohid Yaqub a sizeable €1.4-million (US\$1.7-million) grant to gather evidence on the role of serendipity in science. Yaqub argues that he has found a way to do so.

First, he defines serendipity in a way that goes beyond happy accidents, by classifying it into four basic types (O. Yaqub *Res. Policy* **47**, 169–179; 2018). The first type is where research in one domain leads to a discovery in another — such as when 1943 investigations into the cause of a mustard-gas explosion led to the idea of using chemotherapy to treat cancer. Another is a completely open hunt that brings about a discovery, such as with Röntgen’s X-rays. Then there are the discoveries made when a sought-for solution is reached by an unexpected path, as with the accidental discovery of how to vulcanize rubber. And some discoveries find a solution to a problem that only later emerges: shatterproof glass for car windscreens was first observed in a dropped laboratory flask.

Starting in the archive of US sociologist Robert K. Merton, Yaqub gathered hundreds of historical examples. After studying these, he says, he has pinned down some of the mechanisms by which serendipity comes about. These include astute observation, errors and “controlled sloppiness” (which lets unexpected events occur while still allowing their source to be traced). He also identifies how the collaborative action of networks of people can generate serendipitous findings.

Yaqub, who works at the University of Sussex in Brighton, UK, is now looking to build a team to use this classification system as a framework for mining the world’s scientific grants. By following the publications and patents that emerge from grants, he hopes to find out

how often serendipity arises, and to understand its significance and nature. The hunt will start in biomedicine, but could grow to examine other disciplines.

This seems like a smart way to start attempting something very difficult. And if it proves possible to test the role of serendipity, researchers should do so. Given the paucity of existing evidence, even weak or partial observations could help policymakers to examine the most efficient way to fund research. That could let them balance different modes of funding, for example, or create the environments that encourage serendipity and permit researchers to capitalize on unexpected results.

“Studies rarely try to quantify how much scientific progress was serendipitous.”

If happy accidents are just as likely to occur through goal-oriented research as they are through blue-skies research — witness Roy

Plunkett’s accidental discovery of non-stick Teflon while looking for non-toxic refrigerants — that could weaken the suggestion that serendipity and research-targeting are necessarily at odds with each other. (The heavy-handed oversight that sometimes goes hand in hand with applied research is a different issue.)

Giving curious minds free rein to explore nature may well be the best way to generate discoveries, but there is no harm in testing that assumption. As taxpayer demands for scrutiny and accountability grow ever louder, just-so stories about Petri dishes and non-stick frying pans, however compelling, no longer make a convincing case. ■

Early warning

A seminal study 50 years ago warned of the demise of the West Antarctic Ice Sheet.

Fifty years ago, many scientists were looking up. In 1968, the Russians sent the first animals to orbit the Moon (including a couple of tortoises), and NASA’s Apollo programme kicked into gear to produce the first views of Earth from space. But in Antarctica, John Mercer was looking down — and he was concerned about what he saw.

That year, the late Mercer, a glaciologist at Ohio State University in Columbus, first warned about the potential for rapid sea-level rise from melting ice caps. His landmark paper drew on fieldwork at the Reedy Glacier, which feeds into West Antarctica’s Ross Sea (J. H. Mercer *Int. Assoc. Sci. Hydrol. Symp.* **79**, 217–225; 1968). Geological evidence from a former lake, located at an altitude of 1,400 metres in the Transantarctic Mountains, suggested that the area was once awash with open water and floating icebergs. Mercer took that as evidence that the entire

West Antarctic Ice Sheet had once melted away.

The paper was an intriguing synthesis of the science of the times. Using multiple lines of evidence, Mercer sought to explain how sea levels could have risen by 6 metres in the previous interglacial period, around 120,000 years ago. The melting of Greenland or the East Antarctic Ice Sheet could not explain it, because both are located on solid earth and would respond relatively slowly to warming. By contrast, much of the West Antarctic Ice Sheet is grounded well below sea level. That makes it a “uniquely vulnerable and unstable body of ice”, Mercer wrote.

Many credit a 1974 paper by Johannes Weertman, a geophysicist at Northwestern University in Evanston, Illinois, with providing a technical explanation for how such a massive ice sheet could disintegrate (J. Weertman *J. Glaciol.* 13, 3–11; 1974). And the late Bob Thomas, a NASA glaciologist, spent years investigating and explaining how floating ice shelves acted as corks, stemming the flow of land-bound glaciers into the sea. But Mercer still deserves credit for sounding the alarm.

It took a while for the idea to take hold. Advanced numerical ice-sheet models developed in the late 1980s tended to downplay the risk of rapid ice loss from western Antarctica, and the Intergovernmental Panel on Climate Change suggested in its 1995 report that Antarctica as a whole was stable. But evidence to the contrary mounted: the massive Larsen A and B ice shelves collapsed in 1995 and 2002, respectively, followed by a major rift in Larsen C in 2017. In 2014, a team of scientists declared that the loss of ice in the Amundsen Sea Embayment had accelerated and appeared “unstoppable”.

The future of the ice sheet, which holds enough water to boost global sea levels by more than three metres, is now at the top of the Antarctic research agenda. Scientists are still scouring the world for palaeoclimate records to pin down past sea-level change, modellers are refining their

calculations and fieldwork continues apace. As early as next month, the US National Science Foundation and the UK National Environmental Research Council are expected to jointly announce the recipients of a US\$25-million fund for research on the future of the Thwaites glacier, which flows into the Amundsen Sea. Satellite measurements indicate that melting there has doubled in the past several years, and now accounts for roughly 10% of the global sea-level rise. In a 1978 paper

“The future of the ice sheet is now at the top of the Antarctic research agenda.”

in *Nature*, Mercer updated his arguments in clear and elegant terms. “A disquieting thought is that if the present highly simplified climatic models are even approximately correct,” he wrote, “this deglaciation may be part of the price that must be paid in order to buy enough time for industrial civilisation to make the changeover from fossil fuels to other

sources of energy” (J. H. Mercer *Nature* 271, 321–325; 1978).

That thought still rings frighteningly true. Thus far, the 2015 Paris climate agreement, which commits the world to limiting warming to 1.5–2 °C, remains intact, despite the objections of US President Donald Trump. But grand commitments aside, the governments of the world, and by extension the citizens that they represent, have yet to demonstrate that they are up to the task of reducing greenhouse-gas emissions quickly enough to avert the most disastrous consequences.

Fifty years is the blink of an eye in geological terms, but it is long enough for science to raise its voice. It might feel like pushing against the tide, but researchers have to keep making the point that strong action on emissions could still prevent the worst. Without it, significant sea-level rise will become a certainty. In the long run, higher oceans could well become one of humanity’s most obvious self-inflicted wounds. ■

Outside interests

Nature Research journals will ask authors to declare non-financial conflicts.

What makes a conflict of interest in science? Definitions differ, but broadly agree on one thing: an influence that can cloud a researcher’s objectivity. For some people, that influence can be money. But there are other influences that can interfere, such as institutional loyalty, personal beliefs and ambition.

Nature and the other Nature Research journals (including the *Nature* research and reviews journals, *Nature Communications*, *Scientific Reports*, *Scientific Data*, the *Nature Partner Journals* and the *Communications journals*) are taking into account some of these non-financial sources of possible tension and conflict. From February, authors of research articles, reviews, commentaries and research analyses will be asked (and expected) to disclose them (see go.nature.com/2ddg12z).

For this purpose, competing interests (both financial and non-financial) are defined as a secondary interest that could directly undermine, or be perceived to undermine, the objectivity, integrity and value of a publication through a potential influence on the judgements and actions of authors with regard to objective data presentation, analysis and interpretation. Non-financial competing interests can include a range of personal and/or professional relationships with organizations and individuals, including membership of governmental, non-governmental, advocacy or lobbying organizations, or serving as an expert witness.

We recognize that not everybody shares the same level of concern about non-financial conflicts. Some argue, for example, that because non-financial conflicts cannot be removed, whereas financial conflicts can, focusing on the former could send a message that it’s enough to simply declare financial conflicts rather than remove them. And few would agree with the judge in Scotland who, in a 2005 case, concluded that

non-paid expert witnesses were more likely to be biased (because they wanted to push an agenda) than the highly remunerated experts who spoke on behalf of a tobacco company (L. Friedman and R. Daynard *Tob. Control* 16, 293; 2007).

Numerous studies have demonstrated that financial competing interests in industry-sponsored research have the potential to introduce bias into study design, analysis and reporting; by comparison, the impact of non-financial competing interests has been much less well studied. Nevertheless, it is fair to expect that these associations could colour study design, interpretation and the subsequent reception of published findings; to guard against that, a number of clinical and biomedical journals have required disclosures of non-financial interests for several years. At a time when there is increasing scrutiny of the scientific process, transparent disclosures that allow readers to form their own conclusions about the published work are the best way to maintain public trust.

Nature journals will make full disclosure statements available to peer reviewers as part of the review process and will publish them online. However, although we will facilitate disclosure during the peer review and publication process, the responsibility for appropriately disclosing, managing and eliminating competing interests rests with the authors and their institutions. If we become aware of undisclosed interests that could qualify as a competing interest, in most cases we will amend the published work by issuing a correction. However, in rare cases in which the competing interest is important enough to raise concerns about the reliability of the study, more-serious action may be warranted. Nature Research journals already invite peer reviewers to exclude themselves in cases in which there is a significant conflict of interest, financial or otherwise. And journal editorial staff are required to declare to their employer any interests.

The Nature journals’ competing financial interest policy for authors, which was first introduced in 2001, focused on primary research articles only. We expanded the remit in subsequent years to include review articles and other types of externally authored material, including News & Views, book reviews and opinion articles. The current move is the latest in an evolving process, and we welcome feedback on the change. ■



Take peer pressure out of peer review

Until we study the social dynamics of review panels, assessments will be suboptimal, explains Gemma Derrick.

In its most recent round of university assessments, the United Kingdom tried something new. To judge the value of research beyond academia, review panels in 2014 included a much greater proportion of non-academics than in the previous assessment. For example, pharmaceutical scientists evaluated output in clinical medicine, and government infrastructure experts sat on engineering panels. Despite the shake-up, the university rankings changed little.

That was probably not because the experts within and outside academia agreed on what makes for research that has real-world value. Instead, it seems that the non-academics had little influence.

I interviewed reviewers before and after the evaluations — I study the culture of knowledge production. Those of all backgrounds told me that the interactions on panels were very much academically led. The non-academics had trouble penetrating what one described as “quite a strong culture”. Academics acknowledged and accepted that outsiders were sidelined: their value was in validating the evaluation. One called their presence “a bit of tokenism”; another said that it “provided a type of political capital”.

The UK education council is now assembling reviewers for its next assessment in 2021. It is mainly concerned with getting a mix of academics and non-academics onto panels, and selecting which industries to represent. Without a strategy for how members will work together, these are meaningless efforts.

I think that reviewers of different stripes are not genuinely reaching consensus. There are too few practices to help them do so, and too little knowledge is available to develop these tools. Much work has been done on how to get experts to come to better decisions, but it is unclear how well it applies to confidential reviewing panels.

If peer review is to work as intended (and as commonly assumed), we need to make sure that diverse perspectives are considered amid consistently cliquey groups of academics. In other words, before funding agencies shove a group of strangers into a room and insist they deliver a decision within a strict time limit, we need a better understanding of how these panels actually function.

My own and others' observations show that a peer-review panel is not like some collaborative mural, where everyone contributes a piece to the picture. It is more like a tug of war — with a rope that has many ends. Evaluators form alliances and join various ends of the rope. This sets the panel's dominant mode for dictating how all proposals are assessed. Those outside this framework are quickly silenced, even if they were recruited for their perspective.

The situation undermines what peer review is supposed to accomplish. Peer review is esteemed because, unlike assessments based on metrics, it can incorporate human judgement: panellists

are charged with considering how well a project fits particular goals or with accounting for mitigating circumstances, such as illness, in researchers' productivity.

The system rests on the assumption that experts will work together to air, debate and consider varied views. A sustained collective effort is expected to manage conflicts, to catch weaknesses and mistakes, and to make sound judgements about how to spend public money. These presumed qualities provide political legitimacy, and the outcomes of academic evaluation are accepted by the wider community.

These benefits accrue only if the process is perceived to be fair and informed. Meanwhile, troublingly reductive metric-based evaluations threaten to dominate, with performance defined by strictly measurable formulae. These cost less and can be touted as more objective.

I think the better investment is in learning how to evaluate and improve human review. Unfortunately, efforts to assess review are often thwarted. Most studies of panels can consider only the inputs and outputs, without understanding why some proposals find favour, but not others. Because confidentiality is so prized, getting access to panels took me more than a year. This is at odds with the drive for science and for government decision-making to become more accountable. Worse, it stifles efforts to improve. The study of review panels is essential to optimize the process and to demonstrate that optimal review is valued.

Even limited observations are yielding preliminary pointers that can themselves be evaluated. For example, the Swedish Research Council recently suggested that assigned seating could keep panel members who are already well known to each other from sitting together, and so encourage participation by women and international members.

Many other ideas are worth trying. Pre-evaluation training could help panellists understand how, consciously or otherwise, they might silence competing ideas. Splitting panels into experts who evaluate proposals in isolation and others who make decisions based on blinded assessments could reduce groupthink. Last, peer-review panels could include a non-academic chair to encourage debate from all and actively challenge the consensus.

We should test these strategies with quasi-experimental simulations and by directly observing more panels in action. To ensure the future of peer review, we must understand how to do it better. ■

Gemma Derrick co-directs the Centre for Higher Education, Research and Evaluation at Lancaster University, UK. Her book, *The Evaluators' Eye* is published this month.
e-mail: g.derrick@lancaster.ac.uk

A PEER-REVIEW
PANEL IS NOT
LIKE SOME
**COLLABORATIVE
MURAL;**
IT IS MORE LIKE A
TUG OF WAR.

SEVEN DAYS

The news in brief

SPACE

Challenge cancelled

The Google Lunar XPRIZE's US\$30-million purse will go unclaimed, the XPRIZE Foundation announced on 23 January. Intended to reward the first private team to land a rover on the Moon, the prize has had its deadline pushed back multiple times since it was first announced in 2007. Five teams were racing to meet a 31 March cut-off, but they were stymied by "fundraising, technical and regulatory challenges", the foundation said in a statement. Some of the teams have stated that they will continue their activities with outside funding. The challenge might continue with a new sponsor or with no cash reward, the foundation said.

Falcon Heavy test

Private spaceflight company SpaceX tested all 27 of the engines on its Falcon Heavy rocket at Cape Canaveral, Florida, on 24 January. The roughly ten-second test sent enormous clouds of steam billowing above the rocket, punctuated by the popping roar of dozens of engines. The Falcon Heavy has three boosters — a step up from the company's workhorse, Falcon 9, which has only one. SpaceX of Hawthorne, California, hopes to launch the Falcon Heavy on a maiden flight in the coming weeks, carrying founder Elon Musk's red Tesla sports car. Eventually, the company wants to use the monster rocket to send astronauts to deep space.

PENSIONS

Strike threat

Academics at 61 British universities are preparing to strike over possible

changes to their pensions. In November 2017, after a pension scheme was revealed to have a deficit of £12.6 billion (US\$17.7 billion), universities proposed transferring about 190,000 faculty members and other staff from pension policies that guarantee a retirement income to ones that make income dependent on investment return. Talks between employers and the University and College Union (UCU) in London, which represents tens of thousands of academics nationwide, ended without agreement on 23 January. In a previous ballot of UCU members, 88% supported strike action, which is set to start on 22 February.

The board for the pension plan, known as the University Superannuation Scheme, must make a final decision on proposals by the end of June.

DRUGS

Antibiotic gap

Only 2 of 28 antibiotics in clinical trials against 'high-priority' pathogens are accompanied by plans to prevent unnecessary use that can breed antimicrobial resistance, the Access to Medicine Foundation said in a report on 23 January. The non-profit organization in Amsterdam ranked how well 30 companies are doing in developing new antibiotics

to fight pathogens that have become harder to treat, while ensuring that the drugs remain affordable and available where they are needed. The London-based drug company GlaxoSmithKline had the most products against priority pathogens in development, as well as plans to curb misuse.

ENVIRONMENT

EPA lawsuit

A science-advocacy group and an academic researcher have jointly sued the US Environmental Protection Agency (EPA) to block a rule preventing scientists with active agency grants from serving on EPA advisory committees.



WIN MCNAMEE/GETTY

Humanity edges closer to Doomsday

The *Bulletin of the Atomic Scientists* advanced its symbolic Doomsday Clock on 25 January to two minutes until midnight. The only other time the clock — a symbolic measure of humanity's risk of self-destruction — came so close to the apocalypse was at the height of the cold war, in 1953. This year's decision

to push the clock's hands closer to midnight stems from growing nuclear threats and unchecked climate dangers, said Rachel Bronson, president and chief executive of the *Bulletin*, at a press conference in Washington DC to announce the move. See go.nature.com/2gug3vl for more.

ROGER TILLBERG/ALAMY

The Union of Concerned Scientists in Cambridge, Massachusetts, and Lianne Sheppard, a biostatistician at the University of Washington in Seattle, filed suit in federal court on 23 January. Sheppard had to step aside from an EPA grant to remain on the agency's clean-air advisory committee. When EPA administrator Scott Pruitt introduced the rule last October, he said it would prevent conflicts of interest; the lawsuit says that it will give private industry disproportionate influence.

EVENTS

Indian protest

More than 3,000 people signed an online petition rejecting comments made by a Indian higher-education minister who questioned the validity of the theory of evolution. On 20 January, Satyapal Singh told reporters at a conference on ancient Hindu texts that Charles Darwin's theory of the evolution of humans "is scientifically wrong". Singh's comments outraged researchers, who quickly launched an online petition asking the minister to retract his claims. On 23 January, Singh's boss Prakash Javadekar, the senior minister for human-resource development, dismissed the comments and said his ministry would not support anti-Darwinism

activities, such as Singh's proposal to change educational curricula. See page 16 for more.

PEOPLE

Salk switch

Neuroscientist Rusty Gage has been appointed interim president of the Salk Institute in La Jolla, California, the institute said on 22 January. Gage's appointment comes after the resignation of Nobel-prizewinning molecular biologist Elizabeth Blackburn on 21 December; she is now the institute's president emerita. The Salk faces three lawsuits filed last year alleging that systemic gender discrimination there resulted in lower pay and fewer promotions for female scientists.

Noted author dies

Acclaimed US science-fiction and fantasy writer Ursula K. Le Guin died on 22 January, aged 88. Le Guin (pictured) was the daughter of two anthropologists, and often incorporated race and gender issues, as well as social-science themes, into her narratives. In her fictional futures — such as the Hainish universe, in which humans have spread to and evolved differently on a number of planets — Le Guin was known for exploring social interactions more than the technological advances typical of science fiction. Le Guin's



Earthsea fantasy series, written over four decades, featured a young wizard apprentice and is often described as a precursor of J. K. Rowling's Harry Potter books.

Science nominee

Former astronaut James Reilly is US President Donald Trump's pick to lead the US Geological Survey, the White House said on 26 January. Reilly, who holds a PhD in geoscience, is a technical adviser for the US Air Force's National Security Space Institute. During his career at NASA, from 1994 to 2008, he logged 853 hours in space and went on five spacewalks. Before joining the agency, he worked as an oil and gas exploration geologist. If confirmed by the Senate, Reilly would be the second person with a science PhD to be nominated by Trump to

lead a major science agency. See go.nature.com/2gs8iqt for more.

French agency head

The French government appointed computer scientist Antoine Petit as president of its main basic-science agency, the CNRS, on 24 January. Petit had taken over from interim CNRS chief Anne Peyroche ahead of schedule, because she was "prevented" from continuing in the post, the ministry for higher education, research and innovation said on 18 January. Peyroche's employer, the French Alternative Energies and Atomic Energy Commission, has launched an investigation into the results of some research papers that she co-wrote. A CNRS official told *Nature* that Peyroche is not responding to media enquiries. Petit, who was previously head of INRIA, France's computer-science agency, says that he has three priorities for the CNRS: "to develop basic research, multidisciplinary and France's influence worldwide."

FUNDING

Bulgarian budget

European science ministers are due to meet in Sofia, Bulgaria, on 2 February to discuss future European Union research policy. The nation, which took over the rotating EU presidency last month, had said that it plans to invest in science to boost prosperity. But its national science funding is in disarray after the EU blocked €150 million (US\$186 million) of an expected €700 million in funding for research and innovation facilities in Bulgaria because the country failed to identify enough qualified scientists to satisfy the EU's demands. The Bulgarian government had cut its national science spending by around 25% in anticipation of the windfall.

➔ NATURE.COM

For daily news updates see:
www.nature.com/news

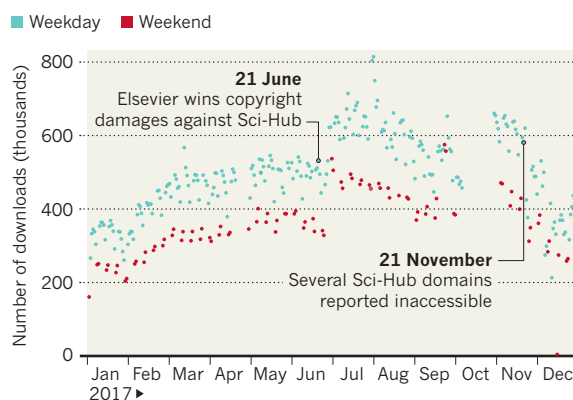
SOURCE: SCI-HUB/BASTIAN GRESHAKE TZOVARAS, LAWRENCE BERKELEY NATIONAL LABORATORY

TREND WATCH

Users of Sci-Hub downloaded more than 150 million papers in 2017, reveal raw data released by the pirate site. The data cover 329 days in 2017 and show that the site's popularity is growing, despite court rulings and efforts to restrict access. An earlier data release covering 6 months in 2015–16 showed around 150,000–200,000 daily downloads, compared with around 460,000 each day last year, according to an analysis by data scientist Bastian Greshake Tzovaras at Lawrence Berkeley National Laboratory in California.

SCI-HUB DOWNLOADS SOAR

The pirate paper site gained its highest-ever traffic figures in 2017, despite court rulings and domain shutdowns.



NEWS IN FOCUS

INDIA Scientists outraged by anti-Darwin comments **p.14**

PALAEoANTHROPOLOGY Fossils in Israel signal earlier human exit from Africa **p.15**

ENVIRONMENT Oil spill in East China Sea presents new type of threat **p.17**



CONSERVATION Owls help farmers kill pests and build partnerships **p.22**

RODGER BOSCH/AFP/GETTY



Bare sand and dead tree trunks stand in a nearly empty reservoir near Cape Town, South Africa.

PUBLIC HEALTH

Cape Town scientists prepare for 'Day Zero'

As water crisis brews, researchers plan to modify studies and prioritize public health.

BY AMY MAXMEN

Panic about the looming water crisis in Cape Town, South Africa, strikes Jodie Miller unpredictably during otherwise ordinary days. "I was making cupcakes this weekend and burned my finger on the pan, and whipped it under the tap," says Miller, a water scientist at Stellenbosch University in Cape Town. "What will we do when no water comes out?"

The city's government estimates that "Day Zero" — when Cape Town becomes the first

major city in the world to run out of water, as reservoirs dip too low to deliver a potable supply — will come on 12 April. The city is experiencing its worst drought in the past century. And with less than 80 days before taps are likely to be shut off for all but essential services, scientists are scrambling to determine how the crisis will affect their research — and their daily lives. Information about the amount of water that universities will be able to draw from municipal or private sources, and for how long, is almost non-existent. Local researchers are also concerned about how the water crisis

will affect their staff and the city as a whole.

"Science needs a functioning ecosystem," says Musa Mhlanga, a cell biologist at the University of Cape Town. "This is very, very serious."

Cape Town is now in its third year of severe drought, which has slashed the supply of surface water on which the city relies. But for many residents, the idea of Day Zero became real only on 18 January, when their mayor announced, "We have reached a point of no return." Until then, the city had hoped that voluntary actions to conserve water would stave off a crisis until winter rains began ►

► and new wells and water-treatment plants came online. But Cape Town's reservoirs have drained faster than predicted: households and businesses have not conserved water as much as the government had hoped during the current dry summer.

According to Cape Town's disaster plan, what water remains in reservoirs on Day Zero — when their levels drop to 13.5% of capacity — will go to hospitals and informal settlements that already rely on communal taps. Most of Cape Town's nearly 4 million residents will have to collect allotments of water from 200 distribution points — a situation that the mayor fears could lead to anarchy.

As recently as 18 January, the University of Cape Town said that, at this stage, it “is not facing the risk of any research activities having to be curtailed as a result of the water crisis”. But researchers there and at other nearby institutions aren't so confident. They're anxious to know whether they might have access to municipal water after the shutdown; although some researchers might be able to import tanks of water from other parts of the country, cost and availability could make that approach difficult to sustain. “The University of Cape Town and its sister institutions are in deep discussions with the city of Cape Town to get clarity” about how much water they will receive, says Valerie Mizrahi, an infectious-disease researcher at the University of Cape Town.

UNDERFUNDED AND UNDERPREPARED

Kevin Winter, who directs the university's Water Task Team, which manages water use across its campuses, lists several reasons why institutions are underprepared for the crisis. Faculty members are just returning to campus after the summer break, as the situation

pivots from hopeful to dire. And although universities have considered water-saving strategies in the past, including water-recycling systems, Winter says that tight budgets have made it difficult to apply those plans. For instance, in recent years, universities responded to student protests demanding free education by agreeing to hold their fees steady despite inflation. “It's not simple to say, give me two to three million dollars towards pipes and pumps and geological surveys,” Winter says. “We are scraping our way through a crisis that might have been averted if we had had the means to do so.”

Now, scientists across Cape Town are drawing up contingency plans for their labs. The

“Science needs a functioning ecosystem. This is very, very serious.”

first step is estimating how much water they will require for basic needs, such as caring for lab animals. Experiments that use less water will be favoured over those requiring more, and at least one biomedical researcher is arranging to move some studies to foreign labs. Many principal investigators are weighing the risks of running experiments that could expose people to fire or chemical injury at a time when water is limited.

Cape Town is a hub for research on HIV and tuberculosis, and medical care associated with those studies will continue. But the water crisis could curtail community outreach on public-health issues. Linda-Gail Bekker, deputy director of the Desmond Tutu HIV Centre, says there would be consequences to interrupting her group's efforts to provide reproductive-health services to young women at high risk of teenage pregnancy and HIV. Bekker worries

that her staff will be unable to work if they must wait in long lines each day to collect water for their households. “I plan to get ahold of the mayor's office this week to see if we can figure something out,” she says.

Robert Wilkinson, director of the Wellcome Centre for Infectious Diseases Research in Africa at the University of Cape Town, spoke to *Nature* from a supermarket where shoppers were snatching up bottled water before it could reach the shelves. Earlier that day, he and his colleagues had discussed how to cut back their clinical trials once Day Zero arrives. Participants would continue to receive medical care, but research would be curtailed; the scientists might request fewer samples of bodily fluids to save the water that would be needed to process them, even for storage.

Wilkinson says that if water is scarce, the urgent priority is to maintain health. “I immediately think about the potential for water-borne illnesses if people — particularly people in poor living conditions — aren't able to maintain personal and institutional hygiene,” Wilkinson says. He also worries about the city's economy, which is reliant on tourism and agriculture.

Then there is Miller, who felt another flash of anxiety recently as she dumped a jug of recycled water on her bare, dirt yard. “It just sat there,” she says. “The water didn't sink in because the land is so parched.” She thought about how droughts and fires can give way to landslides and floods when rain comes. And she considered how pipes can crack when plumbing lies dry for too long. “To be honest, I can't wrap my head around what's happening — a major metropolitan city running out of water,” Miller says. “There are enormous ramifications to this.” ■

FUNDING

Gender bias tilts success of grant applications

But it goes away when reviewers focus on the science.

BY GIORGIA GUGLIELMI

Women lose out when reviewers assess the researcher, rather than the research, according to a study on gender bias. But training reviewers to recognize unconscious biases seems to correct this imbalance, despite previous work suggesting otherwise.

The findings were first posted in December on the bioRxiv preprint server and are currently in review at a journal. They

came out of a 2014 decision by the Canadian Institutes of Health Research (CIHR) to phase out conventional grant programmes, in which reviewers evaluated both the science and the investigator. Instead, the CIHR started one programme that focused its evaluation on the applicants and another that focused mostly on their research. This created a natural experiment that allowed the scientists to analyse the outcomes of nearly 24,000 grant applications and to test whether funding differences were due to the quality of the applicants' research

or to factors related to the applicant, such as gender.

Past studies have looked at gender inequalities in grant funding, but most examined grant programmes that didn't separate their application pool as the CIHR programmes did. Some also didn't consider other factors, such as whether research fields had different ratios of male to female scientists². The new analysis, which took into account applicants' research areas and age — a proxy for career stage — allowed the study authors to draw “more robust conclusions”, says Holly Witteman, a health-informatics researcher at Laval University in Quebec City, Canada, who led the study.

Witteman and her colleagues calculated that, of all the applications submitted to CIHR grant programmes between 2011 and 2016, 15.8% were likely to be successful. And in the conventional grant programmes, the success rate for male applicants was 0.9% higher than the rate for female applicants. When the team analysed the CIHR grant programme that

ISRAEL HERSHKOVITZ, TEL AVIV U focused on the researchers' science, the gap in success rate was the same as in the conventional programmes. But in the grant programme that focused on the applicants' experience and qualifications, the success rate for male applicants was 4% higher than for female applicants. "That's a significant difference," Witteman says.

A RANDOM ACT

However, Witteman warns that the study was not randomized, meaning that there may be differences between male and female applicants, such as their publication records, which might help to account for the different success rates. Her team was unable to account for such factors, because it didn't have access to those data.

"That's a big problem," says Beate Volker, a social scientist at the University of Amsterdam. She says that the CIHR results would reflect bias if they could show that two applicants had similar publication records, but one was preferred over the other. It would be relatively easy to test this by looking at the number and quality of publications for each applicant. But until the researchers do that, the bias is "unproven," Volker says.

Donna Ginther, an economist at the University of Kansas in Lawrence, who analysed racial bias in grant programmes at the US National Institutes of Health³, echoes this concern. But she says it's interesting that the gender differences in funding outcomes disappeared after the CIHR implemented new policies, which included asking reviewers to complete a training module about unconscious bias.

Previous work, Ginther notes, showed that training might stir biases and be counterproductive⁴. The effects of the new CIHR policies suggest the opposite: in the 2016–17 grant cycle, female scientists were as successful as men in both science- and person-focused grant programmes. "It would be helpful to know what kind of training it was," Ginther says.

The CIHR is committed to eliminating bias against women and minorities by educating and evaluating reviewers, says Robyn Tamblyn, scientific director of the CIHR Institute of Health Services and Policy Research in Montreal. "We're just at the beginning," she says.

Witteman now plans to look at the reviewer-training module, to see whether it might help to reduce biases. ■

1. Witteman, H. O., Hendricks, M., Straus, S. & Tannenbaum, C. Preprint at bioRxiv <http://dx.doi.org/10.1101/232868> (2018).
2. Bedi, G., Van Dam, N. T. & Munafo, M. *Lancet* **380**, 474 (2012).
3. Ginther, D. K. *et al. Science* **333**, 1015–1019 (2011).
4. Kaiser, C. R. *et al. J. Pers. Soc. Psychol.* **104**, 5040519 (2013).



An upper jaw and teeth are thought to be the earliest evidence of *Homo sapiens* outside Africa.

PALAEOANTHROPOLOGY

Israeli fossils hint at early migration

Bones suggest humans left Africa 180,000 years ago.

BY EWEN CALLAWAY

The oldest human fossils ever found outside Africa suggest that *Homo sapiens* might have spread to the Arabian Peninsula around 180,000 years ago — much earlier than previously thought. The upper jaw and teeth, found in an Israeli cave and reported in *Science* on 25 January¹, pre-date other human fossils from the same region by at least 50,000 years. But scientists say that it is unclear whether the fossils represent a brief incursion or a more-lasting expansion of the species.

Researchers originally thought that *H. sapiens* emerged in East Africa 200,000 years ago, then moved out to populate the rest of the world. Until discoveries in the past decade countered that story, scientists surmised that a small group left Africa some 60,000 years ago. If so, it would mean that signs of earlier travels were from failed migrations. That evidence includes 80,000–120,000-year-old skulls and other remains from Israel, uncovered in the 1920s and 1930s.

However, recent discoveries have muddled that simple narrative. Some *H. sapiens*-like fossils reported last year from Morocco², which are older than 300,000 years, have raised the

possibility that humans evolved earlier and perhaps elsewhere in Africa. Teeth from southern China³ hint at long-distance migrations some 120,000 years ago. And genome studies have sown more confusion, with some comparisons of global populations pointing to just one human migration from Africa^{4,5}, and others suggesting multiple waves⁶.

EARLY START

In the early 2000s, archaeologist Mina Weinstein-Evron, at the University of Haifa in Israel, and palaeo-anthropologist Israel Hershkovitz, at Tel Aviv University, began a project to excavate a series of Israeli caves. "We called it 'Searching for the Origins of the Earliest Modern Humans'. This was what we were looking for," says Weinstein-Evron.

Their team discovered the jaw fragment in 2002, in Misluya Cave. It is just a few kilometres away from the Skhul cave, one of the sites where the 80,000–120,000-year-old remains were found in the 1920s and ►

"People were coming and going through this land corridor from one continent to another."

► 1930s. Using several different methods, the team estimates the jaw and teeth to be 177,000–194,000 years old.

The remains are unquestionably *H. sapiens*, says team member María Martín-Torres, a palaeoanthropologist at the National Research Centre on Human Evolution in Burgos, Spain. The shapes of the teeth match those of both modern and ancient humans, she says. They also lack features typical of Neanderthals, which lived throughout Eurasia at the time.

The dating seems solid and the fossils are *H. sapiens*, says Huw Groucutt, an archaeologist at the University of Oxford, UK. But he isn't very surprised to see them in Israel. He and his colleagues have previously said that 175,000-year-old stone tools from other sites in the Middle East resemble those used by *H. sapiens* in East Africa⁷.

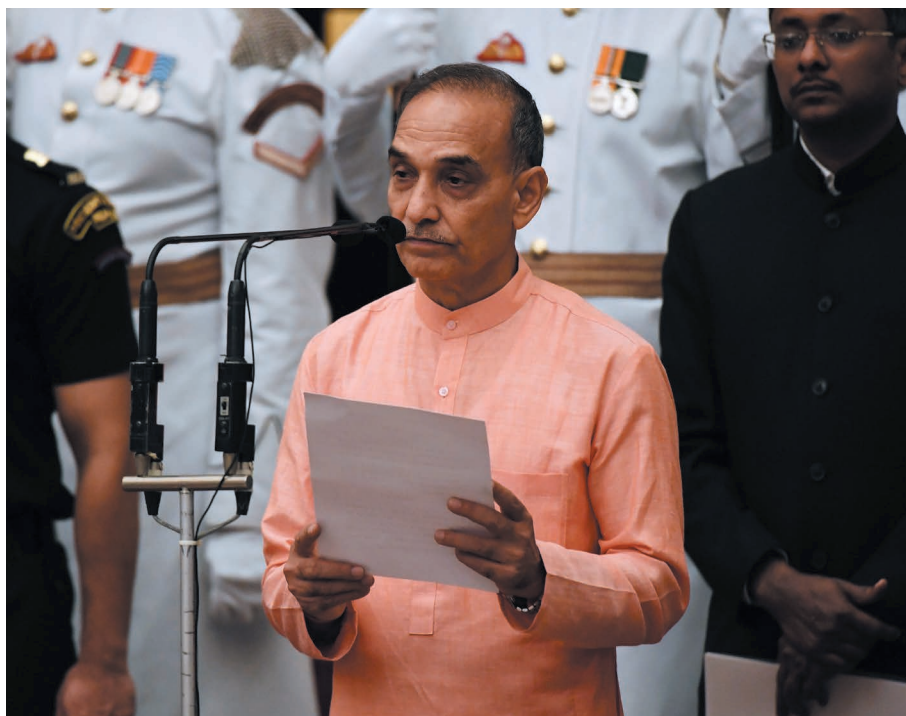
CLOSE ENCOUNTERS

Hershkovitz says that the jaw and teeth point to a long-term occupation of the Middle East by early *H. sapiens*. “It was a central train station. People were coming and going through this land corridor from one continent to another, and it was occupied all the time.” Once in the region, humans probably encountered and interbred with Neanderthals. As evidence, he points to a 2017 ancient-DNA study that suggested interbreeding had occurred before 200,000 years ago⁸.

Wet periods could have drawn humans into the Middle East, but long, dry spells mean that “the region was probably more often a ‘boulevard of broken dreams’ than a stable haven for early humans”, write Chris Stringer and Julia Galway-Witham, palaeoanthropologists at the Natural History Museum in London, in a commentary accompanying the paper⁹.

The fossil could indicate that Israel and the rest of the Arabian Peninsula were part of a larger region in which *H. sapiens* evolved, says John Shea, an archaeologist at Stony Brook University in New York. “We tend to think of Israel as part of Asia for geopolitical reasons, but it is really a transition zone between North Africa and western Asia,” he says. “Plenty of Afro-Arabian animals live there, or did so until recently,” including leopards, lions and zebras. “*Homo sapiens*,” Shea says, “is just another such Afro-Arabian species.” ■

1. Hershkovitz, I. *et al. Science* <http://dx.doi.org/10.1126/aap8369> (2018).
2. Hublin, J.-J. *et al. Nature* **546**, 289–292 (2017).
3. Liu, W. *et al. Nature* **526**, 696–699 (2015).
4. Mallick, S. *et al. Nature* **538**, 201–206 (2016).
5. Malaspinas, A.-S. *et al. Nature* **538**, 207–214 (2016).
6. Pagan, L. *et al. Nature* **538**, 238–242 (2016).
7. Groucutt, H. *et al. Quat. Int.* **382**, 8–30 (2015).
8. Posth, C. *et al. Nature Commun.* **8**, 16046 (2017).
9. Stringer, C. & Galway-Witham, J. *Science* **359**, 389–390 (2018).



Satyapal Singh is a junior minister for human-resource development in India.

INDIA

Anti-Darwin comments outrage researchers

Indian scientists condemn higher-education minister who questioned the theory of evolution.

BY T.V. PADMA

Thousands of scientists in India have signed an online petition protesting against comments by a higher-education minister who last month publicly questioned the scientific validity of Charles Darwin's theory of evolution and called for changes to educational curricula.

The incident continued to simmer when Indian science minister Harsh Vardhan, a medical doctor, declined to comment on his colleague's remarks at a press conference on 24 January. Vardhan said he had not studied Darwin's theory since he was a student and so wasn't qualified to discuss it.

The original comments were made by Satyapal Singh, a junior minister for human-resource development who oversees university education. On 20 January, he told reporters at a conference on ancient Hindu texts in Aurangabad that Darwin's theory of evolution of humans “is scientifically wrong”. Singh added that “nobody, including our ancestors, in written or oral, have said they saw an

ape turning into a man”. Two days later, he proposed holding an international seminar on the subject.

The comments provoked outrage in the Indian scientific community. Vishwesha Guttal, an evolutionary ecologist at the Indian Institute of Science in Bangalore, suggests the remarks are the first time that such anti-evolution opinions have been aired by high-ranking politicians in India. “I have seen these kind of issues (anti-Darwin stance) when I was a student in the US. This was totally unheard of, so far, in India,” says Guttal. “My first thought was, ‘Is this coming to India now?’”

Senior government officials later dismissed the comments. On 23 January, Singh's boss Prakash Javadekar, the senior minister for human-resource development, said that he had asked Singh to refrain from making such remarks. “We should not dilute science,” Javadekar said. He added that his ministry would not support any anti-Darwin activities such as Singh's proposed conference or changing curricula. Singh did not respond to a request for comment from *Nature's* news team.

Scientists reacted swiftly to Singh's comments, launching an online petition asking the minister to retract his claims. Such comments harm the scientific community's efforts to propagate scientific thoughts and rationality through education and research, the petition said, and also diminish the country's image internationally. The petition had collected more than 3,000 signatures when its creators closed it after Javadekar responded to the situation, according to Mukund Thattai, a computational cell biologist at the National Centre for Biological Sciences in Bangalore who signed the petition. "There is strong support for science in India from government departments. But public attitudes can be swayed if people in responsible government positions make such statements," he says.

Soumitro Banerjee, general secretary of the advocacy group the Breakthrough Science Society in Kolkata, thinks that Singh's comments might already have done damage. "The seed of doubt has been planted in the minds of the common people that Darwin's theory of evolution may, after all, be incorrect," says Banerjee, a physicist at the Indian Institute of Science Education and Research in Kolkata.

The minister's comments also prompted a statement from three Indian science academies. "It would be a retrograde step to remove the teaching of the theory of evolution from school and college curricula or to dilute this by offering non-scientific explanations or myths," they said.

Singh's remarks come as India faces a rising tide of pseudoscience. Last year, the Breakthrough Science Society urged researchers to refute unscientific ideas after an astrology workshop was planned at the prestigious Indian Institute of Science in Bangalore. The event was later cancelled.

Vidita Vaidya, a neurobiologist at the Tata Institute of Fundamental Research in Mumbai, says the latest incident highlights the growing gap between the Indian scientific community, policymakers and the public. "It is the responsibility of the scientific community to engage much more actively to ensure that science education and research in this country continue to thrive," she says. ■



Flames envelop the *Sanchi* oil tanker in a picture taken on 13 January.

ENVIRONMENT

Spill in East China Sea raises big questions

Never before has so much light crude oil poured into the ocean.

BY CALLY CARSWELL

When the Iranian oil tanker *Sanchi* collided with a cargo ship, caught fire and sank in the East China Sea in mid-January, an entirely new kind of maritime disaster was born. Two weeks later, basic questions remain unanswered about the size of the spill, its chemical make-up and where it could end up. Without that crucial information, scientists are struggling to predict the incident's short- and long-term ecological consequences.

"This is charting new ground, unfortunately," says Rick Steiner, a former University of Alaska professor in Anchorage who has studied the environmental impacts of oil spills and consulted with governments worldwide on spill response. "This is probably one of the most unique spills ever."

The infamous spills of the past — such as the Deepwater Horizon disaster in the Gulf of Mexico in 2010, or the *Exxon Valdez* tanker rupture in Alaska's Prince William Sound in 1989 — involved heavier crude oil. That oil can remain in the deep ocean for years, and it

has chronic impacts on marine life. The *Sanchi* carried a little more than 111,300 tonnes of natural-gas condensate, a lighter, more volatile petroleum product that doesn't linger as long in the environment. Condensate has never before been unleashed into the sea in large quantities.

Unlike heavy crude, condensate doesn't accumulate in shimmering slicks on the sea surface, which makes it difficult to monitor and contain. Neither does it sink to the ocean floor, as do some heavier constituents of crude over time. Rather, it burns off, evaporates or dissolves into the surface water, where some chemical components can linger for weeks or months.

"Most oil spills have a chronic toxicological effect due to heavy residuals remaining and sinking over time," says Ralph Portier, a marine microbiologist and toxicologist at Louisiana State University in Baton Rouge. "This may be one of the first spills where short-term toxicity is of most concern."

A significant, but unknown, portion of the *Sanchi*'s condensate probably fuelled the fires that followed the collision. In the waters immediately surrounding the tanker, Portier says, ►



TOP NEWS



Animals worldwide stick close to home when humans move in go.nature.com/2bdeInu

MORE NEWS

- Artificial neurons compute faster than the human brain go.nature.com/2ftzrms
- Science behind bars: How a Turkish physicist wrote research papers in prison go.nature.com/2enhe61

NATURE PODCAST



Reframing humans' arrival in India, and the many hazards facing coral reefs nature.com/nature/podcast

► the conflagration and gaseous fumes would have killed off or injured phytoplankton, along with birds, marine mammals and fish that were caught in the vicinity when the tanker ignited.

UNCHARTED TERRITORY

Moving beyond the fire, the impact of the accident becomes harder to discern. That's because the exact chemical composition of the condensate has not yet been made public, Steiner says, and because no one knows how much of the condensate dissolved into the water.

"The part I'm most worried about is the dissolved fraction," Steiner says. Toxic chemicals in the condensate could harm plankton, fish larvae and invertebrate larvae at fairly low concentrations at the sea surface, he says. Fish could suffer reproductive impairments as long as chemicals persist in the water, and birds and marine mammals might experience acute chemical exposure. "In a turbulent, offshore environment, it dilutes fairly quickly," he says. "But it's still toxic."

Because this type of spill is new, Portier says, scientists don't yet understand the ultimate consequences of acute exposure to condensate in the sea, or where it's breaking down and dispersing. "That's really where the science is missing," he says.

Researchers are also scrambling to assess where pollutants from the *Sanchi* could

travel. Groups in both China and the United Kingdom have run ocean-circulation models to predict the oil's journey, and the models agree that much of the pollution is likely to end up in a powerful current known as the Kuroshio, which flows past southeastern Japan and out to the North Pacific. The European models suggest that chemicals from the *Sanchi* could reach the coast of Japan within a month. But the Chinese models indicate that they are unlikely to intrude on Japanese shores at all.

Katya Popova, a modeller with the National Oceanography Centre in Southampton, UK, isn't sure why the models disagree on this point. But, she says, the discrepancy points to the importance of forging international collaborations to increase confidence in model projections during emergencies: "This is something that the oil industry should organize and fund to improve preparedness."

Fangli Qiao, an oceanographer at China's State Oceanic Administration in Qingdao, says his group's models indicate that the pollution's probable path overlaps with Japanese sardine and anchovy fisheries. Still, Popova cautions that the models are imprecise indicators of potential harm to fisheries or coastlines.

"All we're saying is, if something is spilled here at this time, we can give you the most probable distribution," she says. "We don't

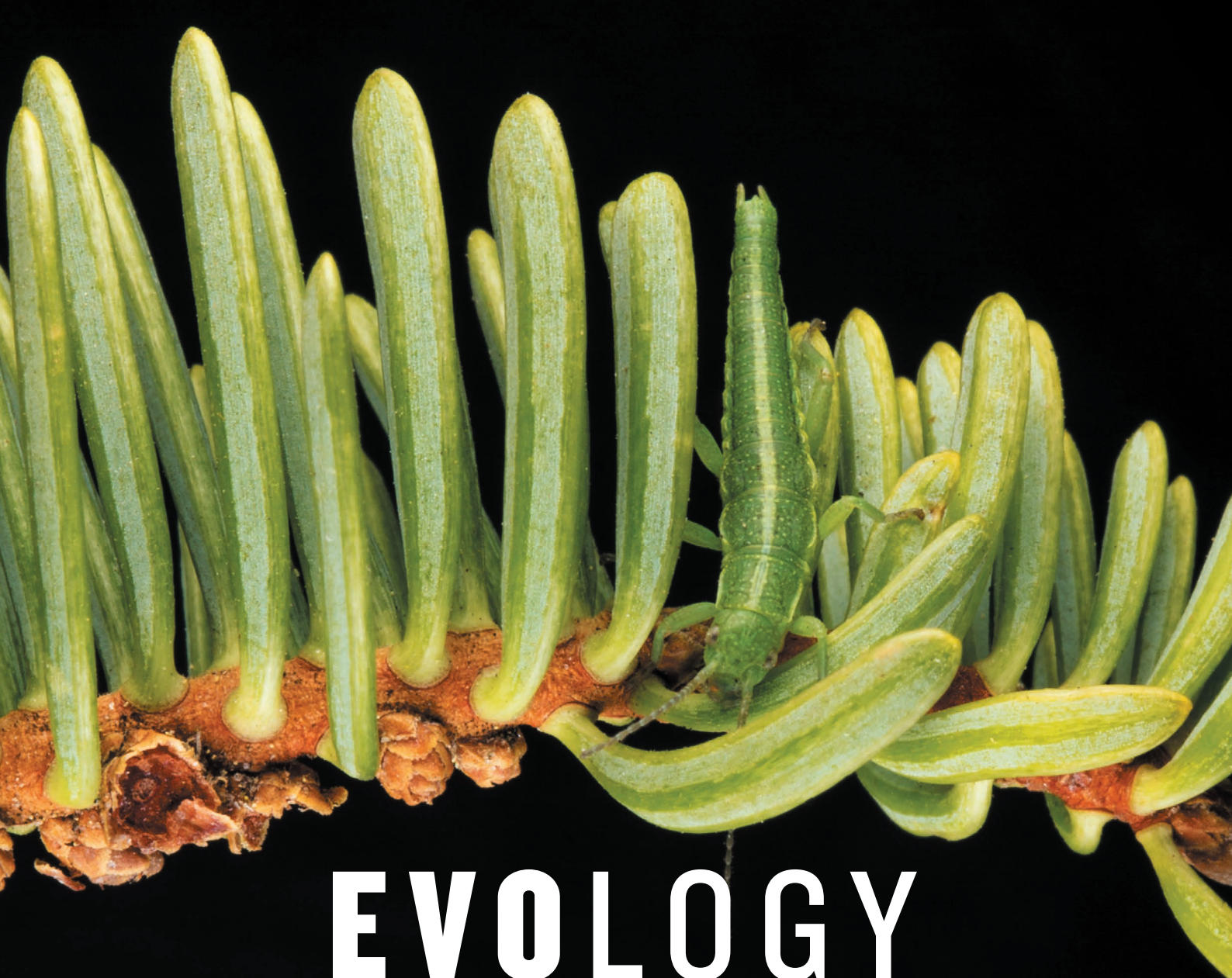
know what type of oil or how much." Those are crucial details because condensate components could degrade or evaporate before reaching important fisheries or shores. "A monitoring programme is the most pressing need right now," Popova says, "to see where it goes and in what concentration."

Yet Steiner says that comprehensive environmental monitoring doesn't seem to have started. Official Chinese-government statements have included results from water-quality monitoring at the wreckage site, but none from the downstream currents that could be dispersing the pollution.

"Time is of the essence, particularly with a volatile substance like condensate," Steiner says. "They needed to immediately be doing plankton monitoring, and monitoring of fish, seabirds. I've seen no reports of any attempt to do that." ■

CORRECTION

The News Feature 'The dark side of light' (*Nature* **553**, 268–270; 2018) erred in saying that differing levels of skyglow had no effect on algae. In fact, it was zooplankton that were analysed. It also cited the wrong journal in reference 9: it should have referred to *Proc. R. Soc. B*.



EVOLUTION

ECOLOGISTS USED TO THINK THAT EVOLUTION WAS TOO SLOW TO AFFECT THEIR STUDIES. THEY WERE WRONG.

BY RACHAEL LALLENSACK

It took Timothy Farkas less than a week to catch and relocate 1,500 stick insects in the Santa Ynez mountains in southern California. His main tool was an actual stick.

“It feels kind of brutish,” says Farkas. “You just pick a stick up off the ground and beat the crap out of a bush.” That low-tech approach dislodged hordes of stick insects that the team easily plucked off the dirt.

On this hillside outside Santa Barbara, there are two kinds of bush that the stick insect (*Timema cristinae*) inhabits. The creature comes in two corresponding colorations: green and striped. Farkas and his fellow ecologists knew that the stick insects had evolved to blend in with their surroundings. But the researchers

wanted to see whether they could turn this relationship around, so that an evolved trait — camouflage — would affect the organism’s ecology.

To find out, the team relocated mixtures of green and striped insects to different plants, so that some insects’ coloration clashed with their new home. Suddenly maladapted, these insects became targets for hungry birds, and that caused a domino effect¹. Birds drawn to bushes with mismatched stick insects stuck around to eat other residents, such as caterpillars and beetles, stripping some plants clean. “That this evolutionary force can cause local extinction is striking,” says Farkas, an ecologist at the University of New Mexico in Albuquerque. “It affects the entire community.” All this happened because of an out-of-place evolutionary trait.

Ecologists have generally ignored evolution when studying their systems; they thought it was impossi-

ble to test whether such a slow process could change ecosystems on observable timescales. But they have come to realize that evolution can happen more quickly than they assumed, and a wave of studies has capitalized on this idea to observe evolution and ecology in unison.

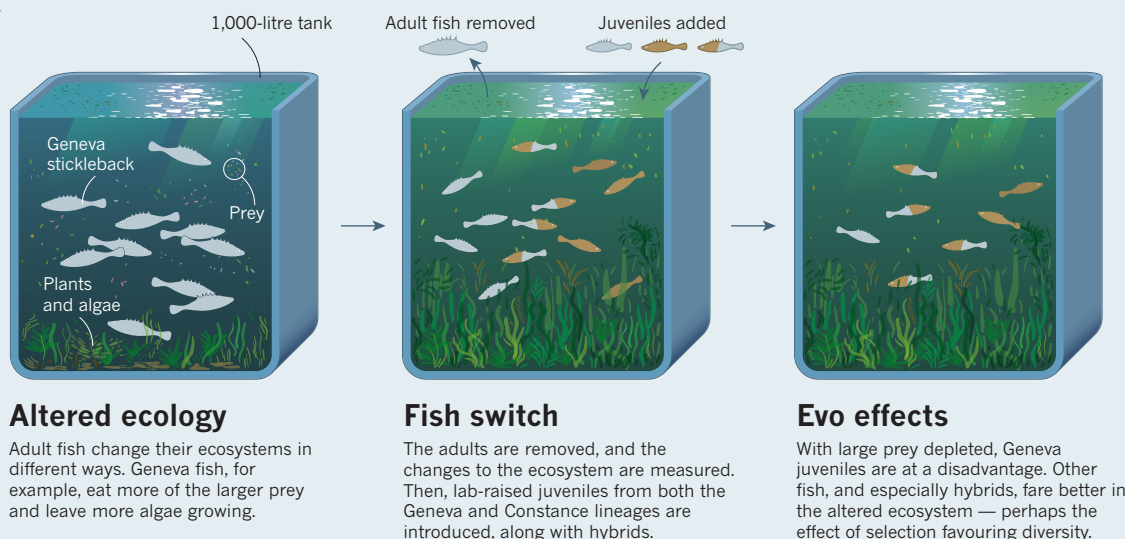
Such eco-evolutionary dynamics could be important for understanding how new populations emerge, or for predicting when one might go extinct. Experiments suggest that evolutionary changes alter some ecosystems

The coloration of stick insects such as this *Timema bartmani* help it to hide, but might also affect the local ecology.

MORITZ MÜSCHICK

FISHY FEEDBACK

Researchers use model ecosystems to test how the evolutionary traits of one generation of a species can affect the environment and, in turn, the next generation. In this example, tanks are stocked with sticklebacks (*Gasterosteus aculeatus*) from either Lake Constance or Lake Geneva.



just as much as shifts in more-conventional ecological elements, such as the amount of light reaching a habitat. “Eco-evolutionary dynamics is the dragon lots of people are chasing right now,” says Troy Simon, an ecologist at the University of Georgia in Athens.

Rapid evolution can sometimes offset some of the detrimental effects of a warming climate and other known drivers of change; in other cases, it can worsen those effects. Even for the most common processes, such as changes in population size or food chains, ecologists must take evolution into consideration, researchers say. “Everybody realized rapid evolution was occurring everywhere,” says evolutionary ecologist Andrew Hendry of McGill University in Montreal, Canada.

DARWIN IN REVERSE

It all goes back to Charles Darwin’s finches. When the naturalist visited Ecuador’s Galapagos Islands in 1835, he documented some variation in the beaks of finches living on different islands and eating different foods. Years after the voyage, he hinted in his *Journal of Researches* that this variation suggested a tight relationship between the birds’ ecology and their evolution.

Darwin never imagined seeing this in action, because he thought that evolution occurs only at the “long lapse of ages”. But by the late 1990s, ecologists had started to realize that evolution could be observed within a few generations of a given species — a timescale that they could work with.

Organisms that live and die quickly provided some of the early data demonstrating how evolution influences ecology. A key study² published in 2003 focused on algae and rotifers, microscopic predators that feed on algae; both species can tick through up to 20 generations in the course of a couple of weeks. The study mixed the organisms together in tanks and showed that when algae evolve rapidly, they throw off normal predator–prey population dynamics.

Usually, the two species play out a cycle between ‘boom’ and ‘bust’. The algal population grows; the rotifers then gobble them up and their own population explodes. When the predators have depleted the algae, their numbers crash. The algae then rebound and the pattern starts again. But when the researchers introduced different algal varieties — seeding some genetic diversity — the algae began to evolve rapidly and the cycle changed completely. The algal population remained elevated for longer, and the rotifers’ own boom was abnormally delayed because the new algae were more resistant to predation.

Similar studies in aphids³ and water fleas⁴ have confirmed that rapid evolution can affect characteristics of populations, such as how fast they grow. These ecological changes can alter future rounds of evolution and selection. Seeing such rapid evolution in action has changed ecologists’ picture of what they thought was a predictable and fundamental ecological process, and showed how important it is to consider evolution when studying how populations interact. “Everything about ecology has to be re-examined in light of the fact that evolution is more important than we thought,” says Stephen Ellner, an ecologist at Cornell University in Ithaca, New York. “This changes everything.”

FAKE LAKES

After these initial lab studies, ecologists started to think bigger. Experiments conducted indoors at small scales can’t reproduce the intricacies of natural ecosystems, so researchers have been testing their ideas in grander, less artificial set-ups.

Working out whether eco-evolutionary dynamics affect the real world is one of the field’s biggest challenges, says Rebecca Best, an evolutionary ecologist at Northern Arizona University in Flagstaff, because so many uncontrollable factors can affect wild ecosystems.

She has found a middle ground by incorporating natural elements into a tightly controlled experiment. At a site overlooking Lake Lucerne in Switzerland, she and her team set up 50 miniature lakes: large plastic tanks each holding 1,000 litres of water, plus a slurry of sediment, plant life, algae, invertebrates and water collected from three lakes — Geneva, Constance and Lucerne. Once these ‘mesocosms’ were settled, with plankton reproducing and plants taking root, the team introduced into each tank one of two genetically distinct lineages of adult threespine sticklebacks (*Gasterosteus aculeatus*): one lineage from Lake Constance and the other from Lake Geneva. A few weeks later, the researchers removed the fish and replaced them with a mixture of lab-raised juveniles from both locations, plus some hybrids of the two lineages.

They found⁵ that how the adults had manipulated their environments affected the survival of the next generation of fish (see ‘Fishy feedback’). If the adult fish removed prey of a certain size, for example, younger fish that shared characteristics with the adults — in this case, mouth size — went hungry. Juveniles that were different from the former occupants fared better. The study showed that the traits of the adult fish shaped the environment for the next generation — enough to dictate the evolutionary trajectory of those that followed.

Best says that her mesocosm experiments are more sophisticated and realistic than lab studies, but less easy to control. Ideally, she says, the team would run the experiment in the field, but that would come with its own obstacles, such as having to factor in the evolution of other species in the ecosystem, or the risk of events such as extreme storms.

Experiments such as Best’s are “vastly easier and more controlled than anything you can do in nature”, Hendry says. But they might not reflect what happens in real ecosystems. “That’s the watershed moment we’re at right now. Does

this actually play out in the real world?”

In the messy real world, it can be difficult to pinpoint the impact of a single feature, either an ecological attribute (such as rainfall) or an evolutionary one (such as a change in camouflage).

A few intrepid ecologists are trying anyway. Last year, a study⁶ on guppies in Trinidad demonstrated that the fish's evolution can drive an ecological change as strongly as an environmental factor: the amount of light available.

The study focused on two populations of guppies (*Poecilia reticulata*) in the northern part of the island. Their habitats differ in several ecological characteristics, including how much shade they receive from the forest canopy, which affects how many algae grow in the streams.

The team moved populations of guppies — which differed in evolved traits such as body proportions and colour — between eight rivers in the watershed, and measured the canopy above the water. In some of the study sites, introducing a new kind of guppy altered algal populations as much as allowing 20% more light to stream onto the water did. Even a natural ecosystem, say the researchers, is a product of evolution as well as ecology.

This experiment did use a more natural setting than many others, but Trinidadian guppies are ecological celebrities that have appeared in hundreds of studies, and the rivers they inhabit have been highly manipulated already. Researchers want to know whether the forces at work in the guppy populations also play out in species that are not necessarily famous for evolutionary dynamics, says McGill ecologist Gregor Fussmann. “We need systems that are generic,” he says.

LIZARD LIMBS

That's exactly what Thomas Schoener, an evolutionary ecologist at the University of California, Davis, and his team have set out to do with two populations of lizard in the Bahamas. Their project is part of an ongoing multigenerational study, begun in 1977. They have been attempting to simulate accelerated evolution by catching curly-tailed lizards (*Leiocephalus carinatus*) and moving them to a string of tiny islands inhabited by brown anoles (*Anolis sagrei*), to see how the ecosystems change as a result.

Curly-tails are natural predators of the smaller brown anole, so when the team first moved the curly-tails onto islands with the anoles, populations of the latter dropped⁷. Spider populations increased when anoles — their main predator — took a hit, and the excess spiders then ate more springtail insects (*Collembola*). Researchers spotted surviving anoles fleeing to the trees to escape their new predator, and that triggered damage to plants. The team knew from previous work⁸ that anoles adapt fairly quickly to tree climbing by favouring shorter-limbed offspring.

But then something unexpected happened. Hurricane Irene hit the islands in

2011, followed by Hurricane Sandy in 2012. Populations of both anoles and curly-tailed lizards crashed. On some islands, anoles were completely wiped out after the storm.

“The hurricanes are a mixed blessing because on the one hand, they give us all kinds of interesting data about disturbance,” Schoener says. “But on the other hand, it can slow down what might be a normal progression of evolution.”

The team has managed to keep its project on track, and is observing evolutionary changes in leg length and the lizards' re-colonization of the islands after the hurricane.

Surprisingly, the anoles that survived the storm have longer limbs than the pre-hurricane population⁷ — the opposite of the team's prediction, but perhaps better for holding on

**“EVERYTHING ABOUT
ECOLOGY HAS TO BE
RE-EXAMINED IN LIGHT OF
THE FACT THAT EVOLUTION
IS MORE IMPORTANT THAN
WE THOUGHT.”**

to branches tightly during a storm. The team has just received funding to study how this evolutionary change will affect the ecosystem.

The hurricanes certainly complicated Schoener's study, but other researchers appreciate the unplanned intervention because it provides a chance to study the consequences of real events and watch the lizards recolonize the islands. Even in the absence of a natural disaster, any number of dynamics could also change the course of an organism's evolution, says Best. “Those potential interactions are going on for everything in the ecosystem.”

She and others say there is plenty more to do, both in the lab and in more-elaborate field studies. Some researchers want to add genetic data to their work, to understand what is driving evolution in the first place. This would tell them whether a particular trait — growth rate, for example — is truly heritable and evolving, rather than a characteristic that can be directly affected by an animal's environment. Genomic data could also help to find hidden characteristics — those harder to observe than body size or growth rate — that might affect ecology.

In a study⁹ of algae and rotifers, Lutz Becks, an evolutionary ecologist at the Max Planck Institute for Evolutionary Biology in Plön, Germany, and his colleagues watched several cycles in which populations waxed and waned as the algae clumped together and dispersed. But when the team looked at individual genes underlying clumping behaviour, they found that their expression varied wildly from one cycle to the next, even though the clumping looked the same. They have since observed co-evolution of three species at once — algae,

rotifers and a virus — and found¹⁰ that the rotifers slowed the rate at which the algae and virus co-evolved. The team plans to repeat this type of experiment, analysing genome data to see how specific details of the algal and viral genes change over time. “We'd like to get to a point where we can actually predict what genomic architecture might be needed for rapid evolution,” says Becks.

Rapid evolution can offset — at least partially — the damaging effects of climate change and other ecological disturbances. In 2011, for instance, a group led by Ellner reanalysed¹¹ 35 years of data from dormant eggs of *Daphnia* water fleas, exhumed from a sediment core in Lake Constance. The data represented periods before, during and after a time when the lake was affected by blooms of cyanobacteria, a microbe with low nutritional value for *Daphnia*. The team found that as the *Daphnia*'s food became less nutritious, juvenile fleas grew poorly and ended up as smaller adults. But after several generations, evolutionary changes caused the growth rate of juveniles to return to normal. And the adults regained some of their lost stature, although they didn't reach the same size as they had before the blooms. The researchers suggest that rapid evolution is likely to occur most often when the environment is changing, but the effects are hidden because they pull in opposite directions. “Evolution is going to be part of how the biosphere responds to climate change,” Ellner says.

Farkas has these questions about evolution and ecology at the front of his mind as he beats the bushes around Santa Barbara and sorts his stick insects. He and his team are planning even more elaborate schemes. They want to catch a full feedback cycle unfolding — ecology affecting evolution affecting ecology once more — all while collecting genetic data. “Comparing how large these effects of evolution will be and understanding when and where evolution is happening is going to be important,” says Farkas. “To me, it's the final frontier. But it's going to take a really long time.” ■

Rachael Lallensack is a journalist based in Washington DC.

1. Farkas, T. E., Mononen, T., Comeault, A. A., Hanski, I. & Nosil, P. *Curr. Biol.* **23**, 1835–1843 (2013).
2. Yoshida, T., Jones, L. E., Ellner, S. P., Fussmann, G. F. & Hairston, N. G. *Jr Nature* **424**, 303–306 (2003).
3. Turcotte, M. M., Reznick, D. N. & Hare, J. D. *Am. Nat.* **181** (Suppl. 1), S46–S57 (2013).
4. Van Doorslaer, W. *et al. Glob. Chang. Biol.* **15**, 3046–3055 (2009).
5. Best, R. J. *et al. Nature Ecol. Evol.* **1**, 1757–1765 (2017).
6. Simon, T. N. *et al. Copeia* **105**, 504–513 (2017).
7. Schoener, T. W., Kolbe, J. J., Leal, M., Losos, J. B. & Spiller, D. A. *Copeia* **105**, 543–549 (2017).
8. Kolbe, J. J., Leal, M., Schoener, T. W., Spiller, D. A. & Losos, J. B. *Science* **335**, 1086–1089 (2012).
9. Becks, L., Ellner, S. P., Jones, L. E. & Hairston, N. G. *Jr Ecol. Lett.* **15**, 492–501 (2012).
10. Frickel, J., Theodosiou, L. & Becks, L. *Proc. Natl Acad. Sci. USA* **114**, 11193–11198 (2017).
11. Ellner, S. P., Geber, M. A. & Hairston, N. G. *Jr Ecol. Lett.* **14**, 603–614 (2011).



Owls for peace

A Middle East programme that uses birds of prey in place of pesticides has inspired cross-border collaboration.

In early 1982, the Tel Aviv University zoo in Israel presented ornithologist Yossi Leshem with an unusual gift: 15 barn owls. The zoo had an owl surplus, and Leshem said he could make use of them. He packed the birds into a van and drove them north to a kibbutz in the Hula Valley. There, farms were plagued so badly with voles that in some years entire fields would echo with their high-pitched squeaking. Leshem, who was working for the Society for the Protection of Nature in Israel, was worried that to control the pests, farmers were overusing a rodent-killing chemical called sodium fluoroacetate, or compound 1080. It had been banned a decade earlier in the United States because of its toxic effects on grizzly bears, hawks and eagles; in Israel, it was killing migrating birds and native egrets. Leshem thought that barn owls (*Tyto alba*), which like to prey on rodents in agricultural fields and are comfortable living close to people, might be the solution. They could control the rodents naturally.

That year, the farmer with whom Leshem planned to work was called up to the Israeli army to fight a war in Lebanon — and was killed. Leshem, who also served in the war, was undeterred. He relaunched his experiment the next year with the help of another farmer, setting up owl nest boxes at a kibbutz called Sde Eliyahu, farther south in Israel's Beit Shean Valley.

More than three decades and numerous conflicts later, the barn-owl

**BY JOSIE
GLAUSIUSZ**

approach to controlling rodents has succeeded beyond Leshem's wildest expectations, spreading across much of Israel and into the neighbouring Palestinian territories and Jordan. And the work has brought together Arab and Israeli scientists at a time of increasing political tensions. "Birds have the power to bring people together, because they know no boundaries," says Leshem, who now also works at Tel Aviv University.

In January, researchers from the Middle East, Mediterranean and North Africa met at a Dead Sea resort in Jordan to see barn-owl nest boxes in the field, discuss scientific findings and hatch plans for similar efforts in Egypt, Cyprus, Greece, Tunisia and Morocco. The programme benefits farmers, biodiversity and sociopolitical networks alike, says conservation biologist Sara Kross at California State University in Sacramento, who in March will host some of the researchers involved in a joint US-Israeli workshop on the topic. Proponents say that the project is more important than ever following US President Donald Trump's announcement in December that the United States would recognize Jerusalem as Israel's capital, which strained relations between Israel and its neighbours. "It's hard to ask people in Jordan, or other Arab countries, to cooperate with Israelis in light of Trump's announcement," says Mansour Abu Rashid, who works with

A Jordanian farmer holds a barn owl at the Sde Eliyahu kibbutz in Israel.

HAGAI AHARON

Leshem and directs the Amman Center for Peace and Development, a non-governmental organization dedicated to promoting dialogue between peoples in the Middle East. “Scientists should continue their cooperation for the benefit and peace of people in the area,” he says.

NATURAL CONTROLS

The owl project did not start smoothly. The first nest boxes, imported from Europe, were not designed for the hot climate, and they baked some baby owls to death. But within about 15 years, the project expanded to the entire Beit Shean Valley, covering about 16,000 hectares, says Shaul Aviel, the Sde Eliyahu farmer who worked with Leshem to start the programme there. The first signs of victory, he says, were in date plantations. Rats can climb and nest in the date palms, and rat-gnawed dates contaminated with droppings cannot be sold. But the damage disappeared when the barn-owl initiative was established. Aviel says that the programme “works 100%” in wheat, dates, olives and pomegranates. But the owls aren’t sufficient to protect all crops; rodents find the green shoots of alfalfa, for instance, irresistible.

Israeli farmers caught on to using the owls, but Leshem realized there was a problem with the programme. Young owls raised in the nest boxes in Beit Shean Valley were spreading through the Jordan Valley, which is shared by the Palestinian territories and Jordan. When they flew over borders, the owls were sometimes poisoned by rodenticides.

But in 2002, Leshem met and began collaborating with Abu Rashid, a retired Jordanian army general who was a key architect of the 1994 Israel–Jordan peace treaty. And by 2008, after some setbacks caused by more political violence in the region, Leshem, Abu Rashid and Imad Atrash, the director of the Palestine Wildlife Society in Beit Sahour, gained funds from the European Union and the US Agency for International Development to launch the project as a cross-border programme.

Researchers set up nest boxes in study sites across three areas of the Jordan Valley, trained farmers and educated local communities about the programme. Now, there are thousands of nesting boxes in Israel and hundreds elsewhere in the region (see ‘Birds across borders’), although curious children have vandalized some of the boxes in the Palestinian territories, Atrash says. Although sceptical of the birds at first — the ghostly white owls are considered a bad omen in many parts of the Middle East — most farmers change their minds after seeing the project’s results, says Abu Rashid. In the Middle East, a pair of barn owls can eat between 2,000 and 6,000 small mammals a year. “The farmers feel it in their yearly production,” he says.

FEWER CHEMICALS

Overall, results are impressive, advocates for the project say. The use of rodenticide goes up and down with the natural rodent cycles of boom and bust, but applications of compound 1080 in Israeli fields have fallen by an average of around 40–60% since the programme began, says Yoav Motro, an ecologist at the Israeli Ministry of Agriculture in Beit Dagan. (Motro’s study has not been published, but he has presented the results at a conference.) Other natural predators of rodents, including kestrels, foxes, jackals and storks, also come in when rodenticide isn’t used, he says. But the best evidence that the barn owls are effective, says Yoram Yom-Tov, a zoologist at Tel Aviv University, is that farmers — who care about their earnings per acre — choose to use them rather than spray chemicals.

Researchers have also learnt about owls’ hunting habits through the programme. Wildlife ecologist Motti Charter of the Shamir Research

Institute at the University of Haifa in Israel has attached radio transmitters to barn owls to show that they can fly 4–7 kilometres away from their nest boxes each night in search of prey — farther than the 500 metres Israeli scientists had originally assumed. (Radio-transmitter research on barn owls in other countries has produced similar findings.) And at the Dead Sea conference this January, ornithologist Alexandre Roulin of the University of Lausanne in Switzerland reported yet-to-be published work showing that the barn owls’ white coloration might enhance their hunting success. Mice are naturally averse to bright light, so they tend to freeze in response to the owls’ ghostly white shine. Roulin, who began collaborating with the project eight years ago after he met Leshem at a scientific meeting, found that on moonlit nights, this effect is enhanced; the Moon’s glow makes the owl’s plumage brighter, making the rodents freeze for longer periods of time.

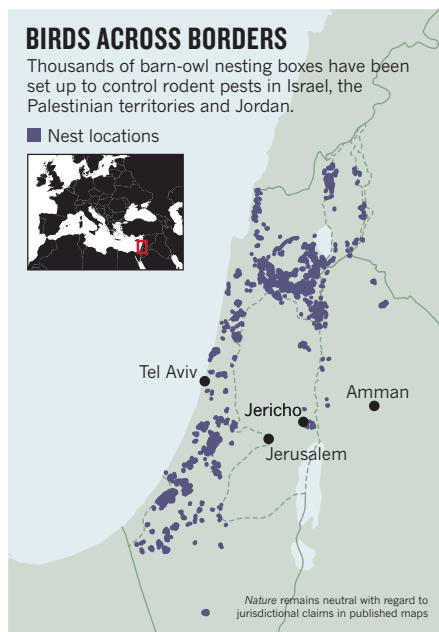
SPREADING THE WORD

Israel, Jordan and the Palestinian territories are not the only places where barn owls have been used to control pests, although they are home to the only cross-border programme. Farmers in Malaysia have used barn owls since 1988 to control rodents in palm-oil plantations — although the government also promotes rodenticide use, says Charter. In California, some farmers have started using barn-owl and kestrel boxes to protect citrus orchards, walnut trees, vineyards and other crops, says Kross. The notion of birds as biological controls dates back to the nineteenth century in the United States, she says, as part of a field called economic ornithology — then defined, according to an 1899 review, as “the study of birds from the standpoint of dollars and cents”. But once pesticides became widespread, the idea died out. Now, the practice is slowly being revived, largely by word of mouth.

Kross and other scientists are watching the Jordan Valley results with interest. “It’s a fantastic project,” says Javier Viñuela, an ecologist at the Hunting Resources Research Institute in Ciudad Real, Spain. In his own country, Viñuela has been collaborating with the non-governmental conservation organization GREFA, which has set up some 2,000 nesting boxes for barn owls and kestrels to control vole numbers. Small-scale experiments inspired by the Israeli experience have started in Argentina and Uruguay, says Charter. And in Cyprus, 27 barn-owl nesting boxes were installed in a pilot collaboration with Israel in 2015; the aim is to install around 60 more this year, says Elena Markitani, a development officer at the non-governmental organization BirdLife Cyprus, which works to conserve wild birds in the country. Martin Hellicar, director of the organization, says that the Middle East programme is a “fantastic thing to try and copy-paste, adapting it to local conditions”. Besides the Israeli project, however, there is a dearth of robust data on the use of owls to control rodents, says Lourens Swanepoel, a conservation ecologist at the University of Venda in Thohoyandou, South Africa.

In general, scientists in the Jordan Valley project avoid talking politics, says Roulin. But they are alert to their work’s political resonance. Last year, Roulin, Charter, Abu Rashid, Leshem and others published a paper entitled “Nature knows no boundaries: the role of nature conservation in peacebuilding” (A. Roulin *et al. Trends Ecol. Evol.* **32**, 305–310; 2017), in which they argue that measures such as the owl project could help to reconcile communities in conflict, building mutual trust without raising the sensitive issues at the root of the strife. “In a conflict area, a project like this or any project in common can help, because the bottom line is the politicians are failing,” says Leshem. “I know I’m not going to solve the problems of the Middle East, but I can do my small part.” ■

Josie Glausiusz is a science journalist in Israel.



► the conflagration and gaseous fumes would have killed off or injured phytoplankton, along with birds, marine mammals and fish that were caught in the vicinity when the tanker ignited.

UNCHARTED TERRITORY

Moving beyond the fire, the impact of the accident becomes harder to discern. That's because the exact chemical composition of the condensate has not yet been made public, Steiner says, and because no one knows how much of the condensate dissolved into the water.

"The part I'm most worried about is the dissolved fraction," Steiner says. Toxic chemicals in the condensate could harm plankton, fish larvae and invertebrate larvae at fairly low concentrations at the sea surface, he says. Fish could suffer reproductive impairments as long as chemicals persist in the water, and birds and marine mammals might experience acute chemical exposure. "In a turbulent, offshore environment, it dilutes fairly quickly," he says. "But it's still toxic."

Because this type of spill is new, Portier says, scientists don't yet understand the ultimate consequences of acute exposure to condensate in the sea, or where it's breaking down and dispersing. "That's really where the science is missing," he says.

Researchers are also scrambling to assess where pollutants from the *Sanchi* could

travel. Groups in both China and the United Kingdom have run ocean-circulation models to predict the oil's journey, and the models agree that much of the pollution is likely to end up in a powerful current known as the Kuroshio, which flows past southeastern Japan and out to the North Pacific. The European models suggest that chemicals from the *Sanchi* could reach the coast of Japan within a month. But the Chinese models indicate that they are unlikely to intrude on Japanese shores at all.

Katya Popova, a modeller with the National Oceanography Centre in Southampton, UK, isn't sure why the models disagree on this point. But, she says, the discrepancy points to the importance of forging international collaborations to increase confidence in model projections during emergencies: "This is something that the oil industry should organize and fund to improve preparedness."

Fangli Qiao, an oceanographer at China's State Oceanic Administration in Qingdao, says his group's models indicate that the pollution's probable path overlaps with Japanese sardine and anchovy fisheries. Still, Popova cautions that the models are imprecise indicators of potential harm to fisheries or coastlines.

"All we're saying is, if something is spilled here at this time, we can give you the most probable distribution," she says. "We don't

know what type of oil or how much." Those are crucial details because condensate components could degrade or evaporate before reaching important fisheries or shores. "A monitoring programme is the most pressing need right now," Popova says, "to see where it goes and in what concentration."

Yet Steiner says that comprehensive environmental monitoring doesn't seem to have started. Official Chinese-government statements have included results from water-quality monitoring at the wreckage site, but none from the downstream currents that could be dispersing the pollution.

"Time is of the essence, particularly with a volatile substance like condensate," Steiner says. "They needed to immediately be doing plankton monitoring, and monitoring of fish, seabirds. I've seen no reports of any attempt to do that." ■

CORRECTION

The News Feature 'The dark side of light' (*Nature* **553**, 268–270; 2018) erred in saying that differing levels of skyglow had no effect on algae. In fact, it was zooplankton that were analysed. It also cited the wrong journal in reference 9: it should have referred to *Proc. R. Soc. B*.

COMMENT

GEOLOGY Long line of triumph and failure — the hunt for Earth's magnetic heart **p.28**

NEUROSCIENCE Antonio Damasio's argument for emotions, appraised **p.30**



PUBLIC HEALTH Study the risk of yellow fever in Asia-Pacific **p.31**

PSYCHIATRY Pamela Sklar, pioneer of mental-health genomics, remembered **p.32**

TAYLOR WEIDMAN/BLOOMBERG/GETTY



Many developing countries, such as Mongolia, have rural economies, so projects that can provide farmers with up-to-date agricultural information are crucial.

Steps to the digital Silk Road

Sharing big data from satellite imagery and other Earth observations across Asia, the Middle East and east Africa is key to sustainability, urges **Guo Huadong**.

The ancient Silk Road trade routes connecting Asia, Europe and Africa lay behind the development of many great civilizations. Today, solar panels and smartphones have replaced silk, and trains and aeroplanes have superseded camels. But the Silk Road spirit of peace, mutual benefit and learning has been revived in an ambitious plan to bridge East and West, launched in 2013 by Chinese President Xi Jinping.

The 'Belt and Road' initiative promises more than US\$1 trillion of Chinese investment in some 60 countries (see 'Belt and Road'). All other nations are welcome to join in. The main aim is socio-economic development through improving the routes for land and sea trade. The initiative will also boost science and technology across the region, for example through research into artificial intelligence, nanotechnology,

quantum computing and smart cities (see go.nature.com/2mvfec6).

But protecting the environment while supporting economic growth will be challenging. The Belt and Road region is home to more than 65% of the world's population. It includes 18 cities that have populations of greater than 10 million, such as Beijing, Cairo, Moscow, Manila and Istanbul.

Environments are diverse and fragile. ►

► Conditions range from the snow, ice and permafrost of the Qinghai–Tibet Plateau to the forests and steppes of Russia and the deserts of Mongolia. Coasts and seas are threatened by rising sea levels, overfishing and pollution. Access to water is a big problem across central Asia. For example, the volume of water in the Aral Sea has shrunk by around 90% in the past 50 years, mainly because the sea and its rivers have been tapped for irrigation.

World Heritage Sites designated by the United Nations Educational, Scientific and Cultural Organization (UNESCO) are endangered by construction, logging, overexploitation and climate change. These include Sumatra's tropical rainforests; Uzbekistan's historic centre of Shakhrisayab; and the world's second-largest raised coral atoll, in the Solomon Islands at the eastern end of Rennell Island.

The economies of many developing countries are rural, with agriculture accounting for more than 25% of gross domestic product. Often, more than 40% of a developing country's workforce is involved in farming. Food supplies can be unreliable.

Natural hazards are another threat. Belt and Road nations experience about 85% of the world's major earthquakes, tsunamis, typhoons, floods, droughts and heatwaves. For example, more than 86,000 people were killed or reported as missing in a massive earthquake in Wenchuan, China, in May 2008. And the 2004 Indian Ocean earthquake and tsunami killed hundreds of thousands of people. Seven of the top ten countries that saw major losses from disasters between 1995 and 2014 are in this region.

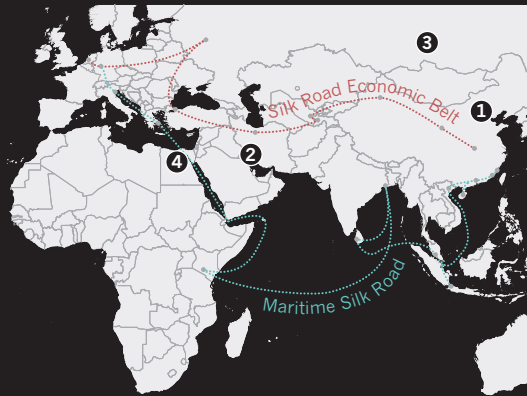
If we do nothing, sensitive environments will be lost and exposure to risks will rise.

To address these problems, a combination of accurate, reliable and timely scientific observations of the state of terrestrial and marine ecosystems is essential — from space, the air and on the ground. However, coverage and infrastructure are poor. Many countries cannot afford to train experts in Earth-observing techniques or install ground stations to monitor soil nutrients or air quality. For example, Kyrgyzstan, Tajikistan, Turkmenistan and Uzbekistan have no Earth-observing satellites or facilities for mass data processing. Local data are rarely shared and are often locked away in government or university archives.

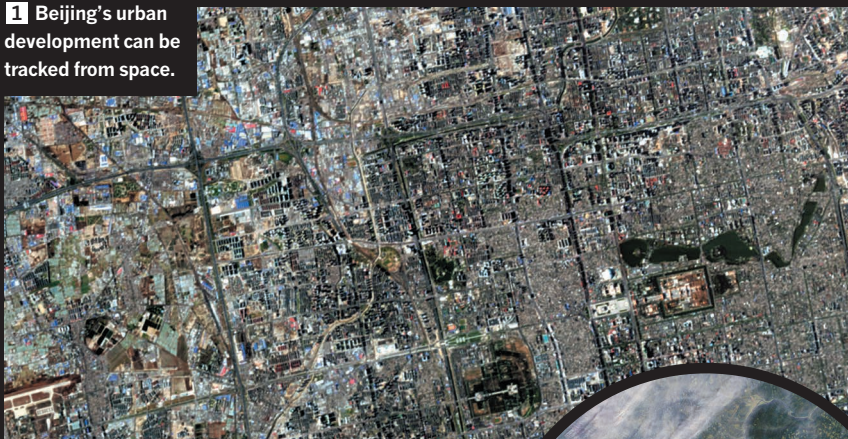
I chair the Digital Belt and Road Program (DBAR) initiated in 2016 by Chinese scientists in cooperation with experts from 19 countries and 7 international organizations. Our aim is to improve environmental monitoring, promote data sharing and support policymaking using big data on Earth observations. The Chinese Academy of Sciences (CAS) is investing more than

Belt and Road

China's US\$1-trillion investment in trade, industry, infrastructure and science is inspired by ancient trade routes that connected East and West (red and blue lines). Shared Earth-observation data are needed to track threats to sustainable development (such as 1, 2 and 4) in more than 60 countries.



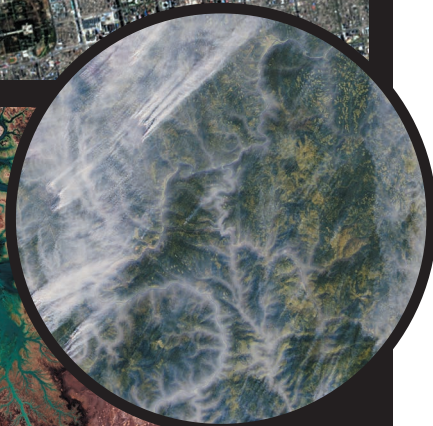
1 Beijing's urban development can be tracked from space.



2 Musa Bay in Iran faces ecological damage from shipping.



3 Wildfires threaten boreal forests in Russia.



4 Development encroaches on the pyramids at Giza.



PHOTOS: EUROPEAN SPACE AGENCY

200 million yuan (US\$32 million) in the next 5 years to support DBAR.

The programme will monitor different types of ecosystem and their evolution, including grasslands, forests, glaciers, urban areas, farmland and coastal regions. Environmental and socio-economic information will be shared through a platform for big Earth data, scheduled for roll-out between 2016 and 2026. This open-access gateway will allow researchers, policymakers and the public to track changes, development and trends. The programme will investigate indices and indicators to feed into the UN's 2030 Sustainable Development Goals.

Here we set out the main scientific challenges and priorities for DBAR. These were presented and discussed in December 2017 at the second DBAR conference in Hong Kong¹.

PROOF OF CONCEPT

There are four main obstacles to a strategy on big Earth data for the Belt and Road region: poor access to data; a digital divide between developed and developing countries; a lack of awareness among some policymakers, local scientists and practitioners of the potential of Earth observations; and too little collaboration. These are long-standing problems — they also slowed emergency responses during and after the Indian Ocean tsunami in 2004, for example.

DBAR's main approach is to work towards a platform that can handle a wide variety of information. Data sets and infrastructure are being assembled, and services should start to become available by the end of 2018. Eight key challenges are being targeted: adapting to climate and environmental change; mitigating disaster risk; managing water supplies; increasing agriculture and food security; protecting natural and cultural heritage; sustainable development of urban areas and infrastructure; managing coasts and marine areas; and understanding changes in high mountains and the Arctic.

For example, in agriculture, the main difficulty faced by most food-insecure countries of the region is a lack of up-to-date information about the supplies, yields and management of crops. DBAR is expanding the cloud-computing-based system CropWatch for monitoring and managing the availability of maize (corn), rice, wheat and soya-bean products. Launched by CAS in 1998, CropWatch provides users from 143 countries or regions with easy access to agricultural information.

For disaster relief and risk reduction, DBAR is developing a platform for sharing Earth-observation imagery. The value of such information in quickly assessing the impacts of extreme events has been proved in China and developed countries, and

needs to be opened to others. For example, following the 2008 Wenchuan earthquake, Chinese rescuers were alerted to 700 people trapped in a village after seeing aerial imagery of “SOS700” written on top of a building.

The processes that shape urbanization need to be understood. Earth observations can reveal trends in the growth of cities and help planners to overcome traffic congestion, energy shortages, urban sprawl and poor basic services. For example, DBAR scientists are modelling the growth of Moscow to inform development in Beijing. The programme is also monitoring the impacts of some big infrastructure projects, including the Mombasa–Nairobi Standard Gauge Railway, Colombo Port City and the Malaysia–China Kuantan Industrial Park.

The entire landscapes of World Heritage Sites — including human influences — need to be protected, not just their monuments². For example, at Angkor in Cambodia, Earth observations that included airborne laser scans revealed the remains of multiple cities aged 900 to 1,400 years old lying beneath the tropical forest floor³. Deforestation and urban sprawl are the main risks that should inform a broader management strategy.

“If we do nothing, sensitive environments will be lost and exposure to risks will rise.”

WAY FORWARD

We plan to focus on five priority areas at DBAR.

Enhance infrastructure. An open platform with shared data, codes and algorithms is urgently needed for analysing the vast amounts of Earth-observation data, which are already daunting and will only increase. The European Space Agency's Sentinel-5P satellite, launched in October 2017, takes 20 million observations of air pollutants and gases each day — 10 times more than previous missions. Cloud computing must therefore be core⁴. It would currently take 1,200 years for one computer to process 3 million planetary-scale satellite scenes; a cloud-computing facility could do it in 45 days⁵. Earth-observing satellite data from upcoming missions will need to be incorporated.

Promote data sharing and interoperability. Data need to be openly exchanged if everyone in the region is to benefit. This will require decisions about suitable formats, information and support for handling them, as well as methodologies and tools to maximize exploitation of the data.

Extend applications to more people. Development across the Belt and Road

region is uneven. To close these gaps, it is necessary to improve common solutions provided by big Earth data⁶. Access to tools such as CropWatch needs to be extended. Use of the digital cloud can allow anyone to access services anywhere across the region, and to accelerate the development of applications for various users.

Identify research opportunities. Knowledge could be discovered within the huge multidisciplinary data sets. For example, studying changes in the land surface of the Yellow River Delta from space over the past 40 years has increased our understanding of how its evolution depends on land use, precipitation and water flows. Researchers must help to raise awareness of the scientific potential and solutions provided by big Earth data, especially in less-developed countries.

Strengthen international collaboration. Belt and Road nations should set up bilateral or multilateral arrangements and stronger links with international scientific programmes and organizations. These include UNESCO, the UN Environment Programme, the UN Office for Disaster Risk Reduction, the Committee on Data for Science and Technology, the Pan-Eurasian Experiment and the Group on Earth Observations.

To help bridge the technical divides between richer and poorer nations, DBAR should set up joint programmes, laboratories and international centres of excellence for gathering experts from participating countries. The programme has already established eight centres of excellence, in Pakistan, Thailand, Finland, Italy, Russia, Morocco, Zambia and the United States.

DBAR has embarked on an ambitious journey to build a digital Silk Road for sustainable development — we invite even more natural and social scientists to join this shared endeavour. ■

*Guo Huadong is chair of DBAR; a professor of remote-sensing science at the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China; president of the International Society for Digital Earth; and former president of the ICSU Committee on Data for Science and Technology.
e-mail: hduo@radi.ac.cn*

1. Digital Belt and Road Program (DBAR). *DBAR Science Plan: An International Science Program for Sustainable Development of the Belt and Road Region Using Big Earth Data* (DBAR, 2017); available at <http://go.nature.com/2evoxcj>
2. Chen, F. L. et al. *Sci. Adv.* **3**, e1601284 (2017).
3. Evans, D. J. *Archaeol. Sci.* **74**, 164–175 (2016).
4. Hansen, M. C. et al. *Science* **342**, 850–853 (2013).
5. Pekel, J.-F., Cottam, A., Gorelick, N. & Belward, A. S. *Nature* **540**, 418–422 (2016).
6. Guo, H. *Big Earth Data* **1**, 4–20 (2017).

The Northern Lights — seen over Iceland — are caused by charged particles in the solar wind hitting Earth's magnetosphere.



GEOPHYSICS

Flip side of geomagnetism

Peter Olson savours a study of Earth's protective force and those who discovered it.

The year 1906 was a momentous one for physics and Earth science. Physicist J. J. Thompson was awarded a Nobel prize for his discovery of the electron; Albert Einstein's two papers on special relativity were being read across the globe; the geologist Richard Oldham discovered Earth's metallic core using seismic waves. And, in a remote location in central France, an obscure physicist called Bernard Brunhes discovered that about 780,000 years ago, Earth's magnetic field had flipped its polarity. The north magnetic pole replaced the south, and vice versa, an event recorded in volcanic rocks across the globe.

In *The Spinning Magnet*, science journalist Alanna Mitchell weaves together the story of Earth's magnetism. She covers its intellectual roots in ancient Greek natural philosophy: Thales of Miletus and Aristotle both speculated on the origins of magnetism. She then follows the hints of it that emerged through the era of classical physics in the nineteenth and early twentieth centuries, and into the



The Spinning Magnet: The Electromagnetic Force That Created the Modern World — and Could Destroy It
ALANNA MITCHELL
Dutton: 2018.

polarity reversals of the geomagnetic field have happened thousands of times in the geological past. We are overdue another. Indeed, Earth's dipole has decreased in strength by nearly 10% since it was first measured by Carl Friedrich Gauss in 1840.

present day, using Brunhes's startling discovery as the subplot. Mitchell adds a generous helping of tension to the mix, thanks to the social implications of that discovery: the potential loss of Earth's dipole, the 'magnetic shield' that deflects harmful particles from the solar wind.

Throughout *The Spinning Magnet*, Mitchell constantly reminds us that

That continues: we are on course to lose our magnetic shield within a couple of millennia. Mitchell even pinpoints where this might already be happening, notably beneath the South Atlantic Ocean. There, on the surface of Earth's liquid outer core, geoscientists are finding evidence that the magnetic field is reversing its direction over several million square kilometres, an area that grows with each passing decade.

So, will the north magnetic pole soon become the south, or is this just a transient episode to be followed by a rebound in magnetic-field strength? Meanwhile, what about our vulnerable infrastructure? How, for instance, will the patchwork US electrical grid respond to extreme solar storms when the magnetic shield goes down? To address these questions, Mitchell has interviewed dozens of geoscientists, space scientists and biologists, offering a readable account of what is probably in store for us, magnetically speaking, and how we got to this point.

ARCTIC-IMAGES/GETTY

More intriguing, however, are the historical players Mitchell describes in this scientific saga. The earliest to get prime billing is Petrus Peregrinus de Maricourt, a medieval French scholar and crusader whose 1269 *Epistle on the Magnet* launched the inquiry into Earth's magnetism in the West. (The subject was already known in the East, especially China, from around 200 BC onwards.) Then comes a trio of Englishmen, from the sixteenth to the eighteenth centuries. These were William Gilbert, astronomer, physician to Elizabeth I and author of the 1600 treatise *De Magnete*; Henry Gellibrand, who showed that Earth's magnetic field varies with time; and Edmond Halley of comet fame, whose representation of the geomagnetic variation became the familiar contour map.

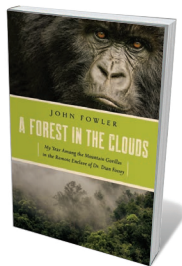
Next are the pioneers of electromagnetism from the eighteenth to the nineteenth centuries: Benjamin Franklin, André-Marie Ampère, Hans Christian Ørsted, Michael Faraday and James Clerk Maxwell. The most prominent woman on the marquee is geophysicist Inge Lehmann, who in 1936 discovered Earth's solid inner core. Finally, there is the large group of other twentieth-century geoscientists who quantified the theory of plate tectonics by using magnetic reversals recorded in igneous and sedimentary rocks. Through this, they demonstrated that the age of the ocean crust increases systematically with distance from the mid-ocean ridges.

Mitchell's portrait gallery is researched with a depth and breadth that make its protagonists' triumphs and failures compelling. She also gives entertaining accounts of today's working geoscientists. They include geologist Jacques Kornprobst, custodian of Bruhnes's legacy, and Daniel Baker, an authority on extreme space-weather events. Her interviews provide insights into their thoughts and actions that transcend the stereotypes of inscrutable nerd or heroic explorer.

The perceptive reader will notice a few disconnects. For one, Mitchell gives short shrift to the explanation of just how Earth generates and maintains its magnetic shield in the metallic core. The human drama behind recent developments, which featured innovative numerical experiments using massive computers, is also curiously omitted. And lastly, the subtitle's dire warnings of impending doom (*The Electromagnetic Force that Created the Modern World — and Could Destroy It*) is an unnecessary distraction, being no match for the real-life stories inside. ■

Peter Olson is in the Department of Earth and Planetary Sciences at the University of New Mexico in Albuquerque.
e-mail: peterleeolson@unm.edu

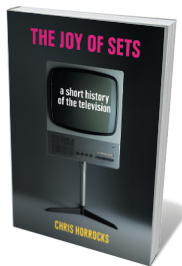
Books in brief



A Forest in the Clouds

John Fowler PEGASUS (2018)

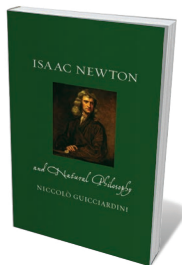
In the early 1980s, John Fowler — then a zoology undergraduate — worked with mountain gorillas at Karisoke, a research centre in Rwanda founded by primatologist Dian Fossey. His book, the only first-hand account of life inside the camp, is both a visceral ethological record and a disturbing portrait of an anguished and embittered Fossey. Framing her unsolved murder in 1985 as that of a scientist-soldier at the front, Fowler ultimately gives Fossey her due as the researcher who taught the world to love a kindred species, even as she became increasingly estranged from her own.



The Joy of Sets

Chris Horrocks REAKTION (2018)

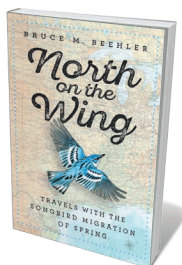
Television, reveals cultural historian Chris Horrocks in this compact chronicle, has tangled roots. The scientific advances of inventor John Logie Baird, broadcast pioneer Paul Nipkow and Karl Braun, inventor of the cathode-ray tube, are only part of the story. A slew of Victorian novels featured visual portals conquering time and space, such as the 'varzeo' in Ismar Thiessen's *The Diothas* (1883). Along with sets, from Baird's 1928 'Noah's Ark' television to today's ultra-thin screens, Horrocks examines the technology's military uses, the ethical furore over content, and its uses as a symbol in art, film and literature.



Isaac Newton and Natural Philosophy

Niccolò Guicciardini REAKTION (2018)

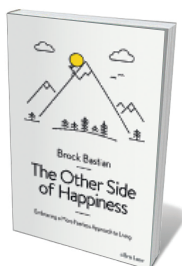
This pithy, nuanced biography of Isaac Newton examines the whole man, as a scientist born into the tumultuous seventeenth century and as an icon (and puzzle) through time. Science historian Niccolò Guicciardini reveals how Newton's theories first received a mixed reception, then became a kind of "cultural fashion" after his death. In the 1930s, his private archive of recondite and theological investigations muddled the picture further. Now, as Guicciardini shows, we are able to see Newton as a brilliant problem-solver eager to crack complexities — in mathematics, metaphysics and alchemy.



North on the Wing: Travels with the Songbird Migration of Spring

Bruce M. Beehler SMITHSONIAN BOOKS (2018)

Many thrill to the spring arrival of avian migrants. Ornithologist Bruce Beehler decided not to wait. Inspired by Edwin Way Teale's 1951 US natural-history road trip, *North with the Spring*, Beehler set off in 2015 to follow, by car, canoe and bicycle, the migration of neotropical songbirds from Texas to Canada. Beehler's 100-day account is both deeply informed by conservation science and history, and lit by euphoric moments such as seeing roseate spoonbills duelling with "absurd spatulate bills", or a cerulean tide of blue jays in flight over Wisconsin wolf country.



The Other Side of Happiness

Brock Bastian ALLEN LANE (2018)

Depressed by positive thinking? Psychologist Brock Bastian concurs. Many theories on the anatomy of happiness have got it wrong, he argues: real well-being involves embracing pain, from the social to the existential. His deft, evidence-based study reveals how avoiding pain backfires; over-protected children become less resilient; and adversity fosters community. If we take calculated risks, stop muffling our sorrow and eschew instant gratification, he avers, our lives will regain clarity of purpose. [Barbara Kiser](#)

The messy biological basis of culture

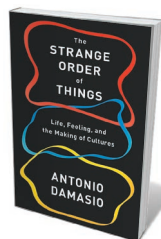
Adrian Woolfson heralds Antonio Damasio's bold argument that emotions define us.

In attempts to define what makes us uniquely human, emotions and feelings are often marginalized. These deeply ingrained, often irrational aspects of our behaviour seem destined to be the poor cousins of the rational cognitive functions that enable the formulation of mathematical theorems or operatic scores. In his bold and important book *The Strange Order of Things*, neuroscientist Antonio Damasio argues that in underestimating the contributions of such 'lower-level' brain phenomena to 'higher-level' cognitive functions, science might have been missing out on some important biology. Similarly, neuroscience's emphasis on the origins of language as a shaper of culture might have eclipsed the role of feelings.

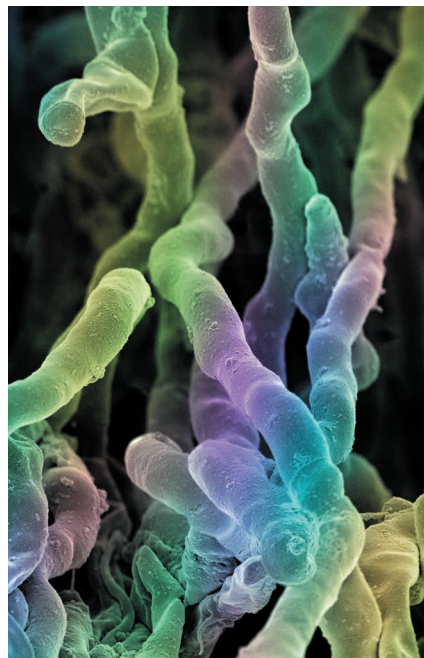
Damasio suggests that our hard-wired drives, urges, compulsions, impulses and automatic responses, such as hunger, desire and pain, originate from "subjective experiences of the momentary state of homeostasis" — that is, the body's routine, humdrum regulation of its visceral function. He argues that there is an organic dialogue between this biological process and the feelings that arise from our continuous scrutiny of it. That diverse penumbra of feelings and impulses, in turn, continuously infuses conscious thought and, ultimately, drives human behaviour.

Human nature is thus distilled from a delicate and protracted negotiation between the beating drums of instinct, shaped by core biological functions, and the attempts of conscious minds to negotiate culturally appropriate outcomes of its mandates. These negotiations are prone to failure, causing dissonance. Damasio theorizes that this generates the profusion of contradictions that both define humans as a species and emerge in our cultural artefacts — from the first stone tools to the Taj Mahal.

Damasio, by unseating the mind from its elevated throne within the brain, delivers an onslaught on one of the core dogmas of conventional neuroscience. In his view, mind is distributed — for instance, to distant anatomical regions such as the peripheral neural networks that control organ function. Thus, different



The Strange Order of Things: Life, Feeling, and the Making of Cultures
ANTONIO DAMASIO
Pantheon: 2018.



Streptomyces coelicolor bacteria (false colour).

tissues in the body contribute incrementally to the mind's function. Damasio's vision offers a new and specific incarnation of the thesis of unified body and mind.

The implications are fascinating. They suggest, for example, that organisms without complex neural structures might possess feelings, and even rudimentary minds. Damasio, incidentally, dates the origin of feelings to the emergence of the first nervous systems, around 600 million years ago.

And at a time when artificial intelligence and machine learning are rapidly advancing and seem poised to disrupt human hegemony, Damasio robustly challenges a belief cherished by transhumanists, who aspire to augment human capacities artificially. That belief is that the myriad aspects of human nature can be captured in their entirety and formalized in the notation of substrate-independent algorithms. A computer program such as Google's AlphaZero can currently master the totality of chess history in around four hours: the time it takes, approximately, to drive from New York City to Washington DC. But, if Damasio is correct, a computer lacking the benefits of biological hardware will never experience the subjective sensation of satisfaction that accompanies a human grandmaster's victory.

Damasio traces core components of

the human "cultural mind", such as social behaviour and cooperation, back to the non-human biology of unicellular organisms present at the inception of life. Bacteria do not sit up at night to contemplate the nature of their existence, and are unable to calculate the trajectories of distant planets. Nevertheless, they are in full command of an impressive repertoire of social behaviours. For example, when nutrients are scarce, bacteria eschew their hermit-like lives and clump together. They can also align into defensive palisades that can confer resistance to antibiotics.

Some of these genetically specified collective behaviours, Damasio argues, may be perceived as the precursors of "moral attitude". For instance, bacteria are able to identify kin through the nature of the molecules they synthesize, and seem to snub "cheaters" that breach social norms. According to Damasio, such behaviours, which provide a basis for social group formation and the emergence of cooperative behaviours, formed the basis for primitive cognition that was later elaborated to generate conscious minds.

Although compelling and refreshingly original, Damasio's thesis would have benefited from a more detailed exposition of the scientific evidence supporting his assertions. The lack of substantial discussion of supporting literature is a significant weakness. As a result, the text reads more like a nineteenth-century philosophical treatise than a contemporary study.

Still, *The Strange Order of Things* addresses several important questions. One of the more compelling consequences of Damasio's organic linkage of mental processes to the properties of biological hardware is that tinkering — such as the use of artificial amino acids and other synthetic-biology technologies — might have unforeseen consequences. Indeed, if Damasio is correct, the precise nature of the substrate that facilitates the performance of biological 'computing machines' will crucially determine their ability to generate mental phenomena. We certainly couldn't expect putative extraterrestrial life forms, which might have evolved by way of unfamiliar biological chemistries and structures, to experience feelings and mental processes in the same way that we do. ■

Adrian Woolfson is the author of *Life Without Genes*.

e-mail: adrianwoolfson@yahoo.com

Correspondence

Yellow fever risk in Asia-Pacific region

The surge in air travel and the adaptive distribution of mosquitoes as a result of urbanization and climate change have accelerated the spread of chikungunya and Zika viruses over the past few years. Yellow fever could similarly spread to the Asia-Pacific region, previously considered off-limits to the virus (see *Nature* **532**, 155–156; 2016). Scientists and public-health officials must take swift pre-emptive action, given that the area has an unvaccinated population of more than two billion people and a limited infrastructure for mounting a response (see go.nature.com/2dtbo6o).

A priority is to collect evidence on the region's mosquito populations, which include *Aedes aegypti*, *Aedes albopictus* and *Aedes scutellaris*. Scientists must study the mosquitoes' competence as vectors for the yellow fever virus, as well as their distribution, biology and susceptibility to insecticides.

Effective surveillance will be crucial. Diagnostic capacity must be improved to distinguish the yellow fever virus from the many other flaviviruses that circulate in Asia. Renewed research into vaccines is also warranted to support a coordinated global response to future outbreaks.

Paul T. Brey Institut Pasteur of Laos, Vientiane, Laos.

Didier Fontenille Institut Pasteur of Cambodia, Phnom Penh, Cambodia.

Hong Tang Institut Pasteur of Shanghai, Chinese Academy of Sciences, Shanghai, China. htang@ips.ac.cn

Don't reject evidence from CFS therapy

We welcome your article on research into chronic fatigue syndrome (CFS), also known as myalgic encephalomyelitis or systemic exertion intolerance disease (see *Nature* **553**, 14–17;

2018). In our view, however, you underestimate the benefits of current treatments: namely, graded exercise and cognitive behaviour therapy (CBT).

As you point out, people with the disease often feel ignored or dismissed by the medical profession — a situation that, as clinicians, we deplore. It is regrettable that some patients and others link behavioural treatments with dismissal of a person's condition, when in fact these therapies can be beneficial. Aside from the results of the PACE trial you mention (co-authored by two of us, T.C. and M.S., among others), further scientific evidence supports the effectiveness of interventions such as exercise and CBT (see, for example, L. Larun *et al.* *Cochrane Database Syst. Rev.* <http://doi.org/cjp6>; 2017 and J. R. Price *et al.* *Cochrane Database Syst. Rev.* <http://doi.org/dcs37d>; 2008).

We think that patients deserve the best research and treatments. In our view, there is no place for stigmatizing any avenue of research or therapy that might help us to improve the lives of people with this long-term debilitating illness.

Michael Sharpe University of Oxford, UK.

Trudie Chalder Guy's, King's and St Thomas's School of Medicine, London, UK.

Jon Stone University of Edinburgh, UK.

michael.sharpe@stx.ox.ac.uk
Competing financial interests declared (see go.nature.com/2at6ng).

Mentoring: a way to help mental health

We contend that the incentive for faculty members to invest in mentoring is limited because of the disempowering structure of academia (see *Nature* **552**, 5; 2017). This leaves graduate students and postdocs with a sense of isolation and without perspective on their accomplishments, both of which contribute to the mental-health

problems experienced by many junior researchers.

Discussion of these mental-health issues has so far focused on strategies for coping with work pressures, for instance by achieving better work–life balance, finding meaning in outreach activities and seeking counselling (see, for example, *Nature* **545**, 375–377; 2017 and *Nature* **539**, 319–321; 2016).

As graduate students, we want instead to shift the discussion to the misplaced incentives in academia that undermine the well-being of both faculty members and students.

Faculty members are expected to excel simultaneously as principal investigators, educators, advisers and administrators. Their job security, however, is determined by the number and perceived impact of papers published and grants secured. This leaves them with little motivation to invest in the mentoring that stands to boost our collective well-being and sustainability.

Rather than place the burden on students to cope with depression, anxiety and burnout, we call on our entire academic community to address the underlying structural problems in academia that cause these mental-health issues in the first place.

Tatyana Perlova* University of Illinois at Urbana-Champaign, USA.

perlova2@illinois.edu

*On behalf of 5 signatories (see go.nature.com/2rau64o for full list).

Mentoring: a rung on the career ladder

As a recipient of a US Presidential Award for Excellence in Science, Mathematics and Engineering Mentoring, I suggest that skills in mentoring should be made a condition for hiring faculty members. Mentoring success should also be included as a criterion for tenure and promotion. This would catapult mentoring to prominence

(*Nature* **552**, 5; 2017).

At the California State University in Northridge, we have shown how well this tactic works. When hiring dozens of faculty members (*Nature* **538**, 171; 2016), we included rigorous evaluations of their teaching and mentoring skills as an essential requirement in addition to their publication records.

The initiative helped to put the university in the top 25 North American institutions classed as 'Rising stars' in 2016 (see go.nature.com/2dfvrb).

Steven B. Oppenheimer California State University, Northridge, California, USA. steven.oppenheimer@csun.edu

Statistics: prevent P-value parrotting

In the same issue in which you propose 'Five ways to fix statistics' (*Nature* **551**, 557–559; 2017), seven of the nine research papers in the biological sciences include *P* values with their figures. In my view, this convention is questionable in most cases, given that the large statistical effects are immediately evident from the respective plots.

To help to fix statistics, I urge authors to instead report a *P* value only when it conveys useful information and is essential for the interpretation of results.

Joachim Goedhart University of Amsterdam, the Netherlands. j.goedhart@uva.nl

CORRECTION

The Outlook article 'The struggle to do no harm' (*Nature* **552**, S74–S75; 2017) mistakenly claimed that the biopharmaceutical firm Cellectis had not responded to requests for interview. A comment from the company is now included in the online version of the story (see go.nature.com/2owayrn).

Pamela Sklar

(1959–2017)

Psychiatrist who sought the genetic roots of mental illness.

When most researchers could not remotely conceive of how to do it, Pamela Sklar's mission was to elucidate the genetic underpinnings of mental illness. A psychiatrist and neuroscientist, she was determined to pursue a systematic analysis of the human genome in search of insights that could help the millions of people globally who battle conditions such as schizophrenia and bipolar disorder.

Sklar was pivotal in pushing forward the idea that many such conditions are polygenic: they result from small effects in many genes, rather than just one or a few. The view — controversial and unproven even 10–15 years ago — transformed psychiatric research. It promoted the understanding that some mental illnesses are no different from complex physical diseases that involve many genes. Sklar's founding and leadership of research consortia were crucial in making the first genetic breakthroughs in these areas. She died on 20 November 2017.

Sklar was born in 1959 in Baltimore, Maryland. During high school, she studied piano at the Peabody Institute in Baltimore, and she enjoyed playing all her life. She earned a bachelor's degree in classics and philosophy from St John's College, a liberal-arts university in Annapolis, Maryland, in 1981.

Sklar showed an early passion for science and medicine. She spent summers working in labs and studying chemistry at Johns Hopkins University in Baltimore; she subsequently earned both her medical degree, in 1985, and her PhD in neuroscience, in 1988, there. Sklar worked in the lab of Solomon Snyder, who revealed the mechanisms of psychotropic and antipsychotic drugs. She did a residency and postdoc work in psychiatry at Columbia University in New York City, with molecular biologist Richard Axel, who in 2004 shared the Nobel Prize in Physiology or Medicine. In 1995, she married molecular biologist and geneticist Andrew Chess, with whom she had two children. She moved to the Massachusetts General Hospital in Boston in 1997.

Sklar was instrumental in founding a psychiatric-genetics programme in 2004 at the then-nascent Broad Institute of MIT and Harvard in Cambridge, Massachusetts. There, in 2007, she helped to launch the Stanley Center for Psychiatric Research. In 2011, she moved to the Icahn School of Medicine at Mount Sinai in New York City and founded the Division of Psychiatric Genomics. It was there that she spent the rest of her career.



Sklar recognized early on the magnitude of the challenge of looking for mental illness's biological roots. Until the past decade, most psychiatric-genetics research used approaches that were successful in the 1980s and 1990s for rare disorders such as Huntington's disease, cystic fibrosis and muscular dystrophies. These techniques sought single genes with strong causal mutations.

By the late 1990s, the lack of success of these efforts in the field of psychiatry revived a long-hypothesized alternative model. Several influential reviews and pilot studies began to explore the idea that some diseases might result from the combined effects of hundreds or thousands of DNA variants, with each genetic variant having a small effect on individual risk. By the mid-2000s, technological advances and the human variation patterns emerging from the Human Genome Project made it possible to explore this hypothesis for the first time.

Sklar realized that the same technologies that were crucial to finding genetic risk factors for diseases such as diabetes could be applied to mental illness. Testing these hypotheses would require huge numbers of willing patients and healthy volunteers. But there was no infrastructure in place for sharing genetic and clinical information globally, and even the idea faced resistance. As a result, Sklar was pivotal in founding what evolved into the Psychiatric Genomics Consortium. She also led and contributed to the first International Schizophrenia Consortium in 2006, and early bipolar-disorder consortia.

In 2009, a seminal publication delivered by these efforts and led by Sklar changed thinking on schizophrenia — and psychiatric diseases as a whole (International Schizophrenia Consortium *Nature* **460**, 748–752; 2009). It reported the first statistically robust evidence that inherited risk variants for schizophrenia were indeed polygenic; the results came from genome-wide association studies. The study also revealed that bipolar disorder and schizophrenia share many genetic risk factors — further supporting other hypotheses about the molecular underpinnings of mental illness.

Critics posited that the lack of strong 'acting' variants suggested that the approach had missed important heritable factors, or that the polygenic patterns were artefacts.

Sklar and her colleagues were undaunted. In 2014, another study definitively established more than 100 distinct risk factors for schizophrenia — cementing the concept (Schizophrenia Working Group of the Psychiatric Genomics Consortium *Nature* **511**, 421–427; 2014). The field is now beginning to learn how to read the many biological clues uncovered, and to pursue them in research into other mental illnesses.

Pamela was an unwavering advocate for the importance of genetics research in improving the lives of those with schizophrenia and bipolar disorder. Throughout her career, she was on the right side of every important decision in our field, whether regarding data sharing, mentorship or in supporting the interests of younger researchers. She led by word and by example — and even through her final months, the passion with which she pursued her research inspired all of us.

She was a fantastic and dedicated mentor to many, and an idol and passionate advocate for women in science. I am proud that Pamela was the first collaborator to welcome me into the psychiatric-genetics community, nearly 20 years ago. Her legacy lives on in the progress we have made, and in the brighter future for people battling mental illness. ■

Mark Daly is chief of the Analytic and Translational Genetics Unit at Massachusetts General Hospital in Boston and co-director of the Medical and Population Genetics Program at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. He was a friend and colleague of Pamela Sklar from 1999.

e-mail: mjdaly@broadinstitute.org

GENOMICS

Regeneration writ large

The assembly of genome sequences for the flatworm *Schmidtea mediterranea* and the salamander *Ambystoma mexicanum* will provide insights into the remarkable regenerative characteristics of these two organisms. **SEE ARTICLES P.50 & P.56**

G. PARKER FLOWERS & CRAIG M. CREWS

Although humans have a limited ability to regenerate after injury, other animals can perform extraordinary regenerative feats. Small flatworms called planarians can regrow their entire bodies, even after being minced into hundreds of pieces. Many species of salamander can regrow whole limbs. For centuries, researchers have sought to understand these abilities, the instructions for which are encoded in DNA. In two papers in *Nature*, Grohme *et al.*¹ and Nowoshilow *et al.*² respectively report the genomes of a planarian (*Schmidtea mediterranea*) and a salamander, the axolotl (*Ambystoma mexicanum*). These studies mark a crucial step towards understanding regeneration.

The genome of *S. mediterranea* is composed of about 800 million bases spread over 4 chromosomes³, which makes it much smaller than the human genome. Even so, until now, the quality of the genome assemblies available for this species has been poor compared with that for other model organisms with larger genomes, such as mice, chickens and zebra fish, owing to problems with large-scale assembly of the planarian genome.

Imagine reconstructing a genome comprising hundreds of millions or even billions of bases from a jumbled library of individual sequence reads, each 500 bases or fewer in length. This is the challenge that has faced researchers using both traditional and next-generation sequencing methods. In theory, if there are enough overlapping individual sequences, a computer algorithm can stitch them together to recreate the genome. But the *S. mediterranea* genome contains abundant repetitive sequences, including virus-like sequences, such as retrotransposons, that have integrated repeatedly and replicated within the genome over the planarian's evolutionary history. These sequences are difficult to distinguish from one another using short samples (Fig. 1). Consequently, previous genome assemblies^{3,4} based on short-read sequencing consisted of more than 100,000 different DNA fragments.

By contrast, Grohme *et al.* (page 56) used a long-read, single-molecule real-time (SMRT) sequencing platform to sequence the *S. mediterranea* genome. Using this strategy,

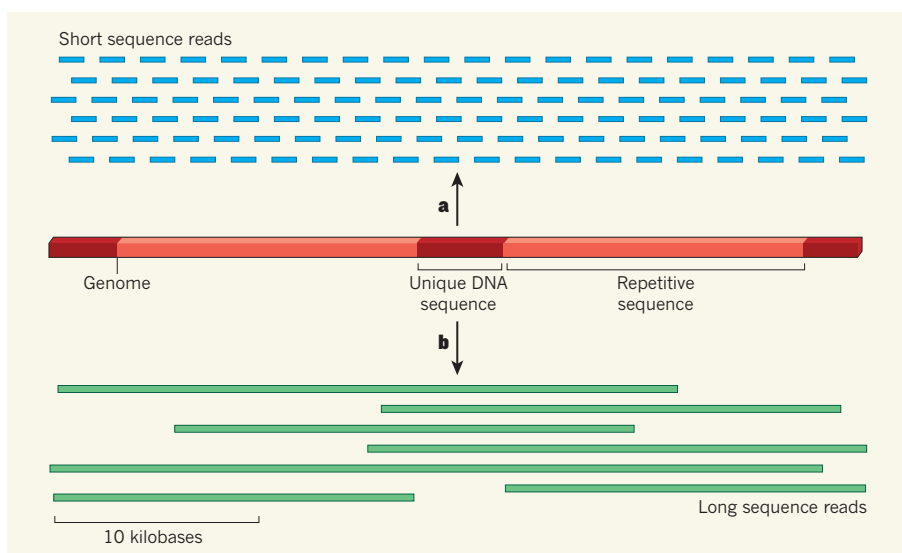


Figure 1 | Assembling repetitive genomes. The genomes of the planarian *Schmidtea mediterranea* and the axolotl *Ambystoma mexicanum* contain many, often long, tracts of repetitive sequences, separated by shorter unique sequences, including those containing protein-coding regions. **a**, Previously, these genomes were sequenced using short-read sequencing methods, which produce many DNA sequences that often cover only repetitive regions, making it difficult to identify their relative positions in the genome. **b**, Two groups^{1,2} have used a method that allows for longer sequence reads (more than 10 kilobases), many of which include both repetitive and unique sequences. This enabled them to assemble genomes for *S. mediterranea* and *A. mexicanum*.

they obtained reads that averaged about 15,000 bases in length — longer than most of the individual fragments from earlier assemblies^{3,4}. They stitched together the sequences by using a computer algorithm called MARVEL, which Grohme *et al.* and Nowoshilow *et al.* developed specifically to improve assembly of long reads from repetitive genomes. This approach bridges many of the assembly gaps caused by repetitive sequences, and produced a genome assembly with fragments more than 1 million bases long on average, which they then ordered into larger scaffolds of around 4 million bases. After this proof of principle, the approach could be applied to a much larger complex genome — that of the axolotl.

Nowoshilow *et al.* (page 50) used SMRT sequencing and MARVEL to assemble the 32-billion-base axolotl genome, which is ten times the size of the human genome. Almost two-thirds of the axolotl genome is made up of repetitive elements, many of which are more than 10,000 bases long. The authors assembled

sequence reads into fragments with a median length of 218,000 bases, which they compiled into scaffolds averaging about 3 million bases. The length and quality of these scaffold sequences are impressive, and demonstrate that some of the most complex genomes can be assembled using this approach.

The group estimates that axolotls have about 23,000 protein-coding genes — slightly more than humans but fewer than found in Grohme and colleagues' planarian assembly. The non-coding sequences within and between the genes are vastly larger than in humans and other vertebrates, mainly because of the expansion of repetitive elements. The group also identified microRNA sequences and genes that are missing in reptiles, birds and mammals, and whose expression was highly enriched in cells of the regenerating limb. Whether any of these candidate genes are crucial players in regeneration will be an interesting subject for future research.

Next, Nowoshilow and colleagues provided

evidence that *Pax3* — a gene essential for development in many animals — is absent in axolotls. They speculate that a related gene, *Pax7*, might compensate for this absence. Indeed, when the authors used gene-editing techniques to inactivate *Pax7*, mutant animals exhibited developmental abnormalities and muscle loss similar to those seen in mice lacking both *Pax3* and *Pax7*.

In a paper recently published in *Nature Communications*, Elewa *et al.*⁵ found that this compensatory action of *Pax7* is probably restricted to a subset of regenerative salamanders. Their genomic study shows that both *Pax3* and *Pax7* are retained in another salamander, the Spanish ribbed newt (*Pleurodeles waltl*), which also has a portfolio of impressive regenerative abilities. Mutational analysis reveals that *Pax7* is not required for normal muscle development and regeneration in this newt, whereas *Pax3* is essential. The ability to investigate and compare the roles of genes in development and regeneration across species in this way marks the beginning of a new era of research in these models of regeneration.

Grohme and colleagues compared the genomes of *S. mediterranea* and other planarian species, and were unable to detect 124 genes that are essential in humans and mice. These genes include some involved in DNA repair and some that have essential roles in protecting against errors in chromosome segregation during cell division. How such organisms manage to thrive without components that have been regarded as essential for life in vertebrates is a fascinating question.

The new genome assemblies, when combined with the sudden ease of genetic manipulation using new genome-editing tools, will make it possible to do experiments that were previously unimaginable in model organisms such as planarians and salamanders. For example, consider the repetitive stretches of DNA that hampered the assembly of the current genomes. In both species, the main contributors to these repeats are retrotransposons. Previous studies^{6,7} have suggested that retrotransposons contribute to important biological processes that shape embryonic development and stem-cell behaviour. Elewa *et al.*⁵ found that retrotransposons are expressed in regenerating limbs in the Spanish ribbed newt, and might in turn regulate gene expression. Whether and how these elements have been co-opted to guide regeneration in various species are among the many exciting avenues of research that can now be explored. ■

G. Parker Flowers and Craig M. Crews are in the Department of Molecular, Cell and Developmental Biology, Yale University, New Haven, Connecticut 06511, USA. C.M.C. is also in the Departments of Chemistry and Pharmacology, Yale University. e-mails: grant.flowers@yale.edu; craig.crews@yale.edu

1. Grohme, M. A. *et al.* *Nature* **554**, 56–61 (2018).
2. Nowoshilow, S. *et al.* *Nature* **554**, 50–55 (2018).
3. Robb, S. M. C., Gotting, K., Ross, E. & Sánchez Alvarado, A. *Genesis* **53**, 535–546 (2015).
4. https://www.ncbi.nlm.nih.gov/assembly/GCA_000691995.1#/st

5. Elewa, A. *et al.* *Nature Commun.* **8**, 2286 (2017).
6. Fort, A. *et al.* *Nature Genet.* **46**, 558–566 (2014).
7. Magiorkinis, G., Katzourakis, A. & Lagiou, P. *Trends Microbiol.* **25**, 876–877 (2017).

This article was published online on 24 January 2018.

BREAST CANCER

A rude awakening from tumour cells

In women who have had breast cancer, drug treatments are often stopped five years after removal of the primary tumour. A meta-analysis shows that these individuals are still at risk of relapse.

GIUSEPPE CURIGLIANO & FATIMA CARDOSO

The main aim of adjuvant therapy, which is given after an apparently successful primary cancer treatment, is to reduce the risk of local and distant metastatic disease relapse owing to residual breast-tumour cells that can persist for years or decades in a dormant state. Our knowledge of the biology of dormant residual disease is cripplingly limited. Writing in the *New England Journal of Medicine*, Pan *et al.*¹ examine rates of metastatic cancer spread in 62,923 women treated for breast cancer and given adjuvant therapy. The authors' findings provide a window on dormancy in this disease.

Pan and colleagues performed a meta-analysis of 88 trials involving women who

had ER-positive breast cancers — subtypes of breast cancer characterized by expression of the oestrogen receptor (ER). The women were all disease-free after five years of scheduled adjuvant endocrine therapy, which involves taking drugs that lower the activity of the ER. The beneficial effects of these treatments for preventing metastasis during the 5 years following diagnosis are not in doubt — instead, the authors' analysis was designed to determine the risk of late metastasis occurring between years 5 and 20, if adjuvant therapy is stopped after 5 years. They found that metastasis occurred at a steady rate for the 15 years after the end of the treatment period.

Remarkably, the most powerful determinants of the risk of recurrence were those originally used to grade the aggressiveness of the primary

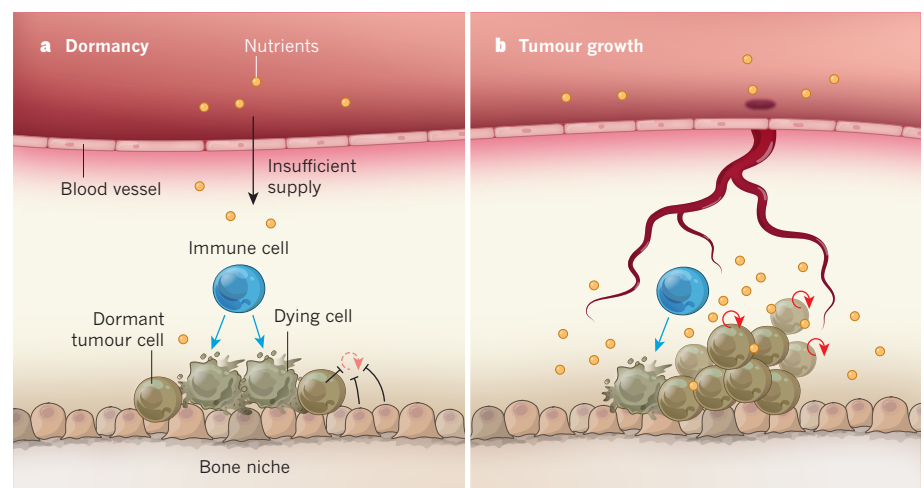


Figure 1 | Multiple mechanisms underlying tumour dormancy. Pan *et al.*¹ report that women who take drugs for five years after their primary breast cancer has been removed are still at risk of a late relapse, because of tumour cells that have migrated to a secondary site such as the bone and adopted a dormant state. **a**, Tumour-cell dormancy can involve many factors, including: poor blood-vessel supply, which means that there is insufficient oxygen and nutrients for proliferation; ongoing surveillance by immune cells that kill some dormant tumour cells; and signals, both tumour-cell-intrinsic and from cells in the surrounding bone niche, that inhibit proliferation (dashed curly arrow). **b**, Relapse occurs when the balance between proliferation and these factors is lost, because of vessel growth, evasion of immune-cell activity or changes that prevent proliferation-blocking signals, leading to tumour growth.

cancer — the diameter of the tumour and the number of lymph nodes containing cancerous cells, which indicates whether the primary disease had spread at the start of treatment. However, even among women with small, node-negative tumours, the risk of metastasis was about 10% over the 15-year period.

Pan *et al.* acknowledge that there are several caveats to their analysis, including a lack of available data on how many women completed their adjuvant treatment, and suboptimal treatment of people who had one particular type of ER-positive breast cancer, dubbed HER2-positive. Nonetheless, it is clear that, even after adjuvant endocrine therapy, women with ER-positive, early-stage breast cancer remain at persistent risk of recurrence for at least 20 years after the original diagnosis. These findings, together with data from another analysis of breast-cancer relapse², might have implications for long-term follow-up strategies, and possibly for treatments.

One way to decrease the risk of relapse might be to extend the duration of adjuvant endocrine therapy — a strategy that is already being tested. We wonder, however, whether this approach will be sufficient to reduce or avoid the risk of late metastasis. One trial³ published in 2016 indicates that increasing the duration of adjuvant therapy from five to ten years significantly improves rates of disease-free survival five years after the treatment ceases, and lowers the incidence of cancer arising in the previously unaffected breast — but does not increase overall survival rates. A second⁴ trial compared continuous adjuvant therapy between years five and ten with an ‘on-off’ therapy that aimed to resensitize cancer cells that might have become resistant to the therapy. It found no difference in rates of metastasis-free survival on completion of either treatment.

Longer follow-ups are needed to better understand the effects of extended adjuvant endocrine therapy, because of the slowly progressive nature of the disease. Nonetheless, it is clear that, although extending the duration of therapy can play a part in preventing late relapses, it might well need to be given for the rest of a woman's life to be effective. This would raise issues of toxicity, compliance and cost.

To best determine other possible ways to reduce the risk of relapse, we must consider what might cause the dormant tumour cells from which metastasis arises to reawaken after many years. Dormant cells have escaped destruction by the body's immune system and entered a microenvironment that supports their survival⁵. Once in this niche, many mechanisms might underlie dormancy — indeed, dormancy can be thought of as a multidimensional status, which involves multiple factors (Fig. 1).

One dimension is cellular dormancy, in which intrinsic or extrinsic factors drive cells into a resting state. Intrinsic factors might

include changes that reduce the cell's drive to divide, similar to those that limit proliferation in cancer stem cells (for example, see ref. 6). These changes could be epigenetic — modulating gene expression without affecting the underlying DNA sequence — or genetic⁵. Extrinsic factors include crosstalk between different cell types in the surrounding microenvironment, such as endothelial cells that line vessels, immune cells and fibroblasts, which make up the structural framework of tissues. A second dimension is vasculature-related dormancy, in which the tumour-cell population is kept small because poor vascularization in the region leads to a lack of nutrients and oxygen. A third dimension is immune-mediated dormancy, in which the immune system controls an expanding tumour-cell population by continuously searching for and eliminating cancerous cells.

If this dormancy status becomes unbalanced owing to changes in any of these factors, dormant cells awaken and metastatic disease develops. Thus, strategies to prevent relapse should aim either to stop dormant cells from awakening or to destroy them when dormant. To achieve this goal, it would be beneficial to identify those patients who will have a late relapse and to tailor a therapeutic strategy to them. ER-positive breast cancers have a low level of intratumoral cell diversity⁷, such as distinct forms of genetic, epigenetic and functional diversity. A better understanding of this diversity might lead to the identification of factors that enable certain cells to become resistant to endocrine therapy and survive in metastatic niches.

Perhaps the major effect of extended adjuvant endocrine therapy is to keep tumour

cells dormant for a longer, but still temporary, period of time. We believe that a combined strategy that concomitantly targets both the tumour cells and the surrounding microenvironment has a higher probability of destroying the dormant cells or of inducing a lifelong state of dormancy, and hence provides a higher probability of a cure. A better understanding of the crosstalk between the dormant cells, their surrounding cell types and the immune system is crucial for developing effective microenvironment-directed therapies. In addition, attention should be paid to the possibility that dormant breast tumour cells undergo evolutionarily conserved programs that lead to a stem-cell-like, prolonged resting state. Maintaining this stem-cell status permanently might be another way to prevent dormant cells from reawakening. ■

Giuseppe Curigliano is in the Department of Hematology and Oncology, European Institute of Oncology, University of Milan, 20141 Milan, Italy. Fatima Cardoso is at the Breast Unit, Champalimaud Clinical Center, Champalimaud Foundation, 1400-038 Lisbon, Portugal.

e-mails: giuseppe.curigliano@ieo.it; fatimacardoso@fundacaochampalimaud.pt

1. Pan, H. *et al.* *N. Engl. J. Med.* **377**, 1836–1846 (2017).
2. Colleoni, M. *et al.* *J. Clin. Oncol.* **34**, 927–935 (2016).
3. Goss, P. E. *et al.* *N. Engl. J. Med.* **375**, 209–219 (2016).
4. Colleoni, M. *et al.* *Lancet Oncol.* **19**, 127–138 (2017).
5. Goss, P. E. & Chambers, A. F. *Nature Rev. Cancer* **10**, 871–877 (2010).
6. Tosoni, D. *et al.* *EMBO Mol. Med.* **9**, 655–671 (2017).
7. Ellis, M. J. & Perou, C. M. *Cancer Discov.* **3**, 27–34 (2013).

This article was published online on 29 January 2018.

ORGANIC CHEMISTRY

Reactive carbon species tamed for synthesis

A highly reactive form of carbon, known as a carbyne, holds great promise for organic synthesis, but has been difficult to prepare. Reactions that produce carbyne equivalents now unleash this synthetic potential. SEE LETTER P.86

ROHAN E. J. BECKWITH

The basis of organic chemistry is the study of carbon-containing compounds with the aim of manipulating carbon atoms to generate new molecules through the formation of carbon–carbon (C–C) bonds. On page 86, Wang *et al.*¹ report a method for harnessing a reactive form of carbon known as a carbyne (Fig. 1a), which has been underused in synthetic chemistry. The findings open the door

to new types of C–C bond-formation reaction.

Conventional approaches to C–C bond formation typically involve well-studied, reactive carbon species. Carbynes, however, have been largely unexplored for C–C bond formation because their high reactivity makes them challenging to prepare and handle. Once formed, carbynes react with each other, with solvent molecules and with other substrates in an uncontrolled manner, producing myriad products. This has limited their applications,

and even efforts to study these species.

Carbynes have previously been formed as complexes with metals, which can then be decomposed to release the free carbyne (see refs 2 and 3, for example). By conducting such decompositions in water at room temperature, researchers have prepared simple compounds that contain C–C triple bonds from the reactions of free carbynes with each other⁴. The formation of undesired side products in these reactions was minimized because the rate of reaction of the free carbynes with water was several times slower than the rate of the carbyne–carbyne reaction⁵. However, product yields were low, and the method has limited applications for synthesis.

Wang *et al.* have devised a clever means of accessing equivalents of carbynes, an approach that has broad synthetic utility. The authors prepared stable precursor compounds that contain two ‘masking’ groups (Fig. 1b). These precursors can be activated by a catalyst in a light-mediated process so that one of the masks is released, generating a carbyne equivalent. Further activation allows the carbyne equivalent to react with substrate molecules, losing the second mask and forming three new bonds in a single synthetic step.

The authors demonstrated the power of this single-step approach by transforming isobutylbenzene (a simple hydrocarbon that contains a benzene ring) into a more-complex system, installing a new ring in a process that forms two new C–C bonds and one carbon–hydrogen (C–H) bond (Fig. 1c). The reaction occurs with high selectivity at particular carbon atoms on isobutylbenzene, and under conditions that chemists would describe as very mild: at ambient temperature and pressure, and using reagents that tolerate the presence of a wide range of groups in the substrate molecule. Such conditions avoid unwanted degradation of the starting materials or the product, thus maximizing the potential product yield.

Wang and colleagues also showed that the unmasking process could be conducted in a stepwise manner by modifying the reaction conditions, allowing the isolation of compounds in which groups known as diazoacetates are attached to benzene rings; one of the masks is retained as part of the diazoacetate group (Fig. 1d). Although methods for making diazoacetates attached to benzene rings have been reported previously, they require a synthetic ‘handle’ — a reactive atom or group — to be present on a benzene ring in the starting material⁶. The authors’ method installs diazoacetates directly at a C–H bond on a benzene ring, and so does not require a synthetic handle. It might, therefore, allow diazoacetates to be attached to complex aromatic systems (compounds that contain a benzene ring, or a related ring system) into which synthetic handles cannot be incorporated.

The authors show that the diazoacetate-installing reaction works effectively for a

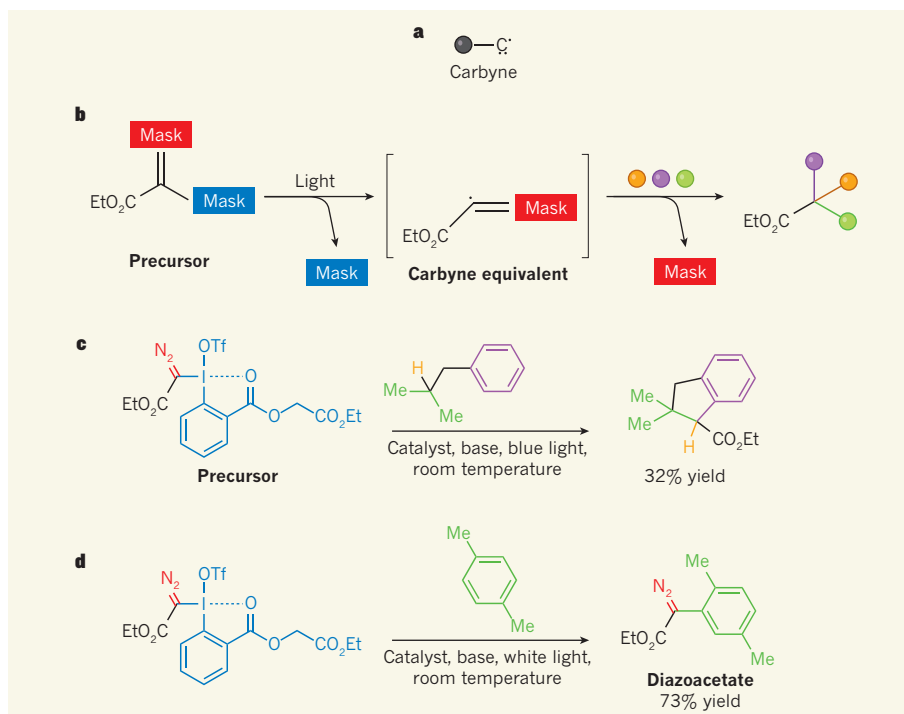


Figure 1 | Carbyne equivalents for synthetic chemistry. **a**, Carbynes are reactive carbon-containing species with potential uses in organic synthetic chemistry. Dots indicate electrons; black sphere represents any organic chemical group. **b**, Wang *et al.*¹ report a strategy that allows equivalents of carbynes to be prepared from precursor compounds that contain two ‘masking’ groups. One mask is removed in a light-mediated process, generating the carbyne equivalent; square brackets indicate that the carbyne equivalent is formed transiently. If the equivalent is generated in the presence of substrate molecules (coloured spheres), it reacts to form three new bonds in a single synthetic step, losing the second mask in the process. Et, ethyl group. **c**, In this colour-coded example, the authors’ precursor reacts with three different groups in a single molecule of isobutylbenzene, when irradiated with blue light at room temperature in the presence of a catalyst and a base. OTf, triflate (SO_3CF_3); Me, methyl. **d**, When white light is used, diazoacetate products form in which one of the masks (a diazo group, red) is retained, as in this example.

range of aromatic substrates. In each case, the diazoacetate is positioned highly selectively at a particular site — even for substrates that contain bulky groups, which often erode site selectivity in other types of reaction. Moreover, the reactions tolerate the presence of a variety of chemical groups in the substrates, and give reasonable yields of products.

The diazoacetate group is especially useful for chemical synthesis because it can be converted into many other groups using metal-catalysed reactions (as Wang *et al.* demonstrate for diazoacetates made using their approach). The availability of a method that allows diazoacetate groups to be incorporated site selectively into structurally complex molecules will be particularly useful for researchers working in drug discovery, because it will enable variants of biologically active molecules to be made rapidly from the molecules themselves, rather than in multi-step synthetic routes from simple starting materials. Indeed, Wang *et al.* demonstrate effective, site-selective diazoacetate incorporation into 12 complex drug molecules.

There are some limitations to the new reactions — for example, drug compounds that contain basic amine groups tend not to be amenable to this approach. The yields obtained

for diazoacetate incorporation into drug molecules are often low to moderate (10–58%) compared with the yields observed for the less-complex substrates (mostly 50–99%). Moreover, substrates that bear strongly electron-withdrawing groups on the aromatic ring also seem to reduce yields and site selectivity relative to other substrates. But these are minor issues compared with the benefits of this approach for drug discovery and development.

Wang and colleagues’ work also has more-fundamental implications for organic chemistry. It offers a new means of generating diazo compounds (the family of compounds to which diazoacetates belong) that allows access to types of diazo molecule that were not previously accessible using other chemistries. In addition, because the reactions generate three new bonds, they provide a way to convert simple reagents into structurally complex molecules in a single step. The authors’ reactions are therefore an invaluable addition to the synthetic chemist’s toolbox.

Finally, Wang *et al.* have reported the first evidence that carbynes can be harnessed effectively for practical organic synthesis. The findings therefore open the door to the exploration of new reactivities of carbon — either through modification of the currently reported masked

carbyne equivalents, or through the development of alternative ones. ■

Rohan E. J. Beckwith is at the Novartis Institutes for BioMedical Research, Inc., Cambridge, Massachusetts 02139, USA.

e-mail: rohan.beckwith@novartis.com

1. Wang, Z., Herraiz, A. G., del Hoyo, A. M. & Suero, M. G. *Nature* **554**, 86–91 (2018).
2. Fischer, E. O., Ruhs, A. & Plabst, D. Z. *Naturforsch. B* **32B**, 802–804 (1977).
3. Levy, O., Musa, S. & Bino, A. *Dalton Trans.* **42**,

- 12248–12251 (2013).
4. Bino, A., Ardon, M. & Shirman, E. *Science* **308**, 234–235 (2005).
5. Bogoslavsky, B. et al. *Angew. Chem. Int. Edn* **51**, 90–94 (2012).
6. Ye, F. et al. *J. Am. Chem. Soc.* **137**, 4435–4444 (2015).

MARINE VIROLOGY

A non-tailed twist in the viral tale

Microscopy studies indicate that the most common viruses in the sea lack a tail structure. However, most cultured marine viruses have tails. A family of these elusive non-tailed marine viruses has now been identified. [SEE LETTER P.118](#)

JULIO CESAR IGNACIO-ESPINOZA
& JED A. FUHRMAN

Bacteriophages, the viruses that infect bacteria, are thought to be the most abundant biological entities on Earth¹. It has been estimated that, if lined up end to end, bacteriophages from the oceans alone would cover a distance of 3 million parsecs, past many distant galaxies². Most bacteriophages cultured in the laboratory or represented in DNA-sequence databases have a ‘tail’ structure, which might take the form of a tube with spider-leg-like protrusions (similar in shape to the base of the Apollo Moon lander). However, electron-microscopy analysis of ocean samples indicates that oceanic viruses

are predominantly non-tailed^{3–5}. Therefore, identifying the ‘missing’ non-tailed marine viruses might improve our understanding of how viruses regulate the microbial systems that control a large fraction of global carbon and nitrogen cycling^{2,6}. On page 118, Kauffman et al.⁷ report the discovery of a non-tailed viral family that they suggest might be an important component of the missing viruses, and they explain why this group might have eluded detection until now.

Viruses can be key drivers of the evolution, community composition and mortality of microorganisms, although their major role in the oceans was not recognized until the 1990s^{2,6}. The number of species a virus can infect is a crucial factor that can influence

gene transfer between species and the spread of viral infection. Bacteriophage action is usually studied using models in which a virus infects only one or a small group of closely related host strains or species, because this is the pattern observed for most of the bacteriophages cultured so far. And yet such cultures represent just a small proportion⁸ of the hundreds of known bacterial phyla⁹.

Kauffman and colleagues sought to expand the breadth of marine viruses known. They analysed water samples collected off the coast of Massachusetts on three separate days, and attempted to identify bacteriophages that could infect any of the 1,334 strains of Vibrionaceae bacteria that they had isolated. The Vibrionaceae are an easily cultured bacterial group that contains the cholera-causing pathogen *Vibrio cholerae*, as well as more-benign relatives, such as bioluminescent bacteria that form symbiotic relationships with fishes and squid.

Of all the bacteria that the authors tested, 239 strains became infected with viruses, and Kauffman et al. isolated 241 previously unknown viruses, of which 18 were non-tailed. They named this non-tailed family (Fig. 1) the *Autolykiviridae*, after Autolykos, an elusive thief from Greek mythology who could not be caught. DNA-sequence analysis revealed that autolykiviruses have small genomes (approximately 10 kilobases in length), which diverge enough from those of known viruses to form their own distinct lineage. They differ from other bacteriophages in sequences encoding a specific structural fold in a capsid protein, which forms the outer viral shell. The putative capsid-encoding sequences identified by Kauffman and colleagues were most like those that encode a fold known as a double jelly roll. This fold was previously associated with non-tailed viruses, in contrast to the HK97 fold found in tailed bacteriophages¹⁰.

To investigate the ecological role of autolykiviruses, the authors performed a monumental host-range analysis, testing the ability of their 241 marine-isolated viruses to infect and kill 318 of their marine bacterial strains. They found that the 18 non-tailed autolykiviruses were responsible for a disproportionately high number of bacterial cell deaths. This was mainly because the autolykiviruses had a substantially wider host range than the tailed group, and could infect multiple Vibrionaceae genera, whereas the tailed viruses studied could not. The authors therefore propose that the impact of autolykiviruses on the marine environment might be fundamentally

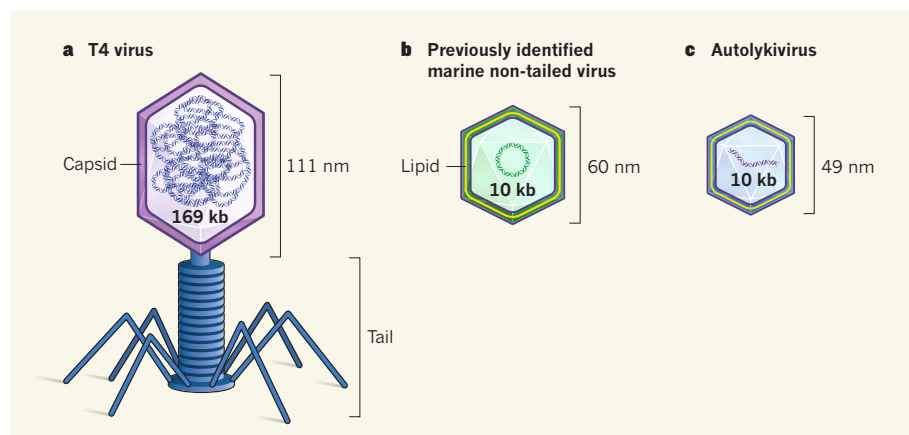


Figure 1 | Different viral forms. Viruses that contain a tail structure are the most common type of bacterium-infecting virus (bacteriophage) cultured in the laboratory or represented in DNA databases. However, in samples taken from marine environments, non-tailed viruses are more common^{3–5}. Kauffman et al.⁷ report a previously unknown family of non-tailed marine viruses. **a**, T4, an example of a tailed virus. Its 169-kilobase genome is enclosed in a capsid structure, made of protein (dark purple), that is 111 nanometres long¹⁵. The average capsid length for tailed marine viruses⁵ is 65 nm. The tail structures in certain other types of tailed virus have a different shape from that of T4. **b**, The cortovirus PM2, one of the few non-tailed marine bacteriophages identified so far. PM2 has lipid (yellow) associated with its capsid. Non-tailed marine viruses have an average capsid size⁵ of 54 nm. **c**, An autolykivirus, a member of a family of non-tailed marine viruses identified by Kauffman et al.⁷. The properties of these bacteriophages are consistent with the presence of lipid.

different from that of tailed bacteriophages. Although this is possible, it is perhaps premature to generalize. Kauffman *et al.* examined only Vibrionaceae. Moreover, some tailed marine bacteriophages have wide host ranges¹¹.

How did these viruses evade previous detection in highly studied systems? The authors note that this could be because some standard viral isolation approaches for lab culture and DNA analysis have sampling biases, which might arise if a virus particle contains lipid, as seems to be the case for the autolykiviruses.

Chloroform treatment is commonly used to disrupt cell membranes as a way of limiting bacterial contamination during bacteriophage isolation; however, Kauffman *et al.* report that chloroform can inactivate autolykiviruses. And when a density-gradient centrifugation approach is used to isolate bacteriophages for DNA-sequence analysis, autolykiviruses are found in a low-density fraction separate from the heavier fraction that contains most bacteriophages (and which is thus usually analysed). An additional snag is the presence of protein covalently bound to autolykiviral DNA that necessitates treatment with a protease enzyme for DNA extraction, an uncommon step in standard large-scale virus-sequencing studies. These technical obstacles should be considered both in future surveys and in the interpretation of existing ones.

Are these autolykiviruses the missing non-tailed marine viruses? Although this discovery certainly illuminates the value of analysing host–virus systems found in marine environments^{11–13}, the *Autolykiviridae* constitutes just a fraction of the missing non-tailed viruses — the authors' quantitative isolation technique demonstrated that only around 7% of the viruses (18 out of 241) cultured on Vibrionaceae were autolykiviruses. Autolykivirus-related sequences are found in DNA sequences of many phyla of bacteria and archaea (another single-celled group that lacks a nucleus). This suggests that other *Autolykiviridae*-like viruses exist, although the authors were unable to estimate the frequency of autolykivirus-like sequences. They observed that the autolykiviral infection cycle was slower than that of tailed bacteriophages in laboratory experiments, which might imply a disproportionately lower effect of autolykiviruses when viruses compete in nature.

Because the autolykiviruses were found by studying hosts representing only a tiny portion of the overall diversity of marine bacteria, the use of hosts from multiple phyla might reveal other viral groups previously missed. Modified experimental protocols might capture those excluded for the same types of technical reason that prevented autolykiviral identification. Finding previously unknown viral groups is crucial, because large-scale DNA surveys across organisms require reference sequences, and this study points us in the right direction. Luckily, a method is available for extracting

viral DNA that avoids the problematic step of density gradients and includes a protease-treatment step¹⁴. This puts the search for further non-tailed viral relatives, and the quantitative study of their effect on marine microbial systems, within reach. ■

Julio Cesar Ignacio-Espinoza and Jed A. Fuhrman are in the Department of Biological Sciences and at the Wrigley Institute for Environmental Studies, University of Southern California, Los Angeles, California 90089, USA.
e-mail: fuhrman@usc.edu

1. Cobián Güemes, A. G. *et al. Annu. Rev. Virol.* **3**, 197–214 (2016).
2. Suttle, C. A. *Nature* **437**, 356–361 (2005).
3. Børstheim, K. Y., Bratbak, G. & Haldal, M. *Appl. Environ. Microbiol.* **56**, 352–356 (1990).

4. Wommack, K. E., Hill, R. T., Kessel, M., Russek-Cohen, E. & Colwell, R. R. *Appl. Environ. Microbiol.* **58**, 2965–2970 (1992).
5. Brum, J. R., Schenck, R. O. & Sullivan, M. B. *ISME J.* **7**, 1738–1751 (2013).
6. Fuhrman, J. A. *Nature* **399**, 541–548 (1999).
7. Kauffman, K. M. *et al. Nature* **554**, 118–122 (2018).
8. Holmfeldt, K. *et al. Proc. Natl Acad. Sci. USA* **110**, 12798–12803 (2013).
9. Hug, L. A. *et al. Nature Microbiol.* **1**, 16048 (2016).
10. Krupovic, M. & Koonin, E. V. *Proc Natl Acad. Sci. USA* **114**, E2401–E2410 (2017).
11. Sullivan, M. B., Waterbury, J. B. & Chisholm, S. W. *Nature* **424**, 1047–1051 (2003).
12. Kang, I., Oh, H.-M., Kang, D. & Cho, J.-C. *Proc. Natl Acad. Sci. USA* **110**, 12343–12348 (2013).
13. Zhao, Y. *et al. Nature* **494**, 357–360 (2013).
14. Steward, G. F. & Culley, A. I. in *Manual of Aquatic Viral Ecology* 154–165 (Assoc. Sci. Limnol. Oceanogr., 2010).
15. Comeau, A. M. & Krisch, H. M. *Mol. Biol. Evol.* **25**, 1321–1332 (2008).

This article was published online on 24 January 2018.

PALAEOCLIMATE

Pollen weighs in on a climate conundrum

Simulations by climate models show that Earth warmed during the Holocene epoch, whereas ocean sedimentary cores suggest that global cooling occurred. An analysis of fossil pollen samples now sides with the models. [SEE LETTER P.92](#)

JEREMY D. SHAKUN

The iconic 'hockey stick' graph shows that global warming of 1 °C during the past century has reversed a 2,000-year-long, modest cooling trend¹. But longer trends are also of interest: how did temperatures change during the rest of the Holocene epoch, which began about 11,700 years ago at the end of the last ice age, and during which human civilization arose? It was previously thought that the Holocene coincided with long-term global cooling², but on page 92, Marsicek *et al.*³ present a reconstruction of Holocene temperatures from North America and Europe that indicate a long-term warming trend. The findings cast light on how well climate models and proxies agree with each other and reproduce the ancient climate, the drivers of climate change during interglacial periods, and the geological context for the current state of the climate.

Global temperature might seem a simple quantity, but its changes during the Holocene are not easy to deduce theoretically. The factors that had the largest effect on climate during this period were gradual variations in the tilt and wobble of Earth's axis. These orbital changes altered the insolation (the amount of sunlight received) for different regions of the world and the seasonal patterns of this

insolation. However, such orbital variations do not change the global mean annual insolation, and thus would not affect global temperature if the climate system responded in a linear way.

But the climate system's response might have been nonlinear — for instance, if responses occurred more strongly in some areas or at some times of the year than others⁴. Indeed, it is well established that orbital forcing gave rise to the glacial–interglacial cycles of the past few million years, with most of the change in global temperature typically being attributed to feedbacks associated with ice sheets and greenhouse gases. Climate models suggest that these two factors would also have dominated global temperature during the Holocene: retreating ice sheets early in the epoch, and rising levels of greenhouse gases later on, both nudged the planet towards warmer temperatures⁵.

The first reconstruction² of Holocene global temperatures to be derived from the geological record was based mainly on sea surface temperatures obtained from the analysis of marine sediment cores. In contrast to the models, this showed that the early Holocene was the warmest part of the epoch, and that global temperature has dropped by about 0.7 °C during the past 5,000 years. The disparity between models and the reconstruction has been dubbed the 'Holocene temperature conundrum'⁵, and two possible explanations



Figure 1 | Pollen grains. Marsicek *et al.*³ have analysed pollen samples preserved in lake sediments from North America and Europe to reconstruct temperatures since the end of the last ice age. The colours used in this micrograph are false.

have been put forward. It could be that proxies of sea surface temperature used in the analysis of the marine cores are biased towards temperatures that occur at certain times of the year, and thus record orbitally driven changes in seasonal temperature, whereas the mean annual temperature changed only little. Or perhaps climate models do not accurately simulate nonlinear feedback responses to seasonal insolation that are stronger at some times of the year than others, and thus yield mean annual temperature changes.

Marsicek *et al.* revisit this problem by producing a new temperature reconstruction based on analyses of pollen grains (Fig. 1) preserved in lake sediments, often referred to as fossil pollen. Pollen-based reconstructions can potentially resolve mean annual and seasonal trends because each plant species has its own sensitivity to winter cold and summer warmth. For example, the populations of some species decline above or below certain temperatures, and so the abundance of pollen from such species in fossil samples can be used as a proxy of temperature.

The authors note that cooling recorded in cores from the North Atlantic Ocean was the main driver of the drop in mean global temperature observed in the previously reported Holocene reconstruction² — in other words, the North Atlantic is the chief source of the conundrum. The authors therefore compiled and statistically analysed hundreds of pollen records from North America and Europe to see whether the cooling trend holds on either side

of the North Atlantic. They find that it does not: the mean annual temperature in their reconstruction rises by 1 °C during the first half of the Holocene, then plateaus, and finally drops slightly during the past 2,000 years. This pattern is encouraging, because it closely matches a previously reported climate-model simulation⁵ for the entire Holocene, as well as the hockey-stick reconstructions of the past two millennia.

Marsicek and colleagues also used this model simulation to show how pollen and marine proxies of surface temperature might record different aspects of the same season, and thus exhibit divergent Holocene trends. They find that cooling in the marine records parallels declining peak summer temperatures in the simulation, with changes in such temperatures being driven by declining summer insolation. Pollen-inferred temperatures, however, increase in step with the amount of heat simulated to have been received at the surface throughout the growing season, which lengthens over time because of rises in greenhouse-gas concentrations and cool-season insolation.

Despite these proposed differences in how the pollen and marine reconstructions record climate over short timescales, the reconstructions exhibit surprisingly similar variability over millennia: surface temperature undergoes a few oscillations of a fraction of a degree Celsius. It is unclear what to make of these fluctuations, but they resemble variability produced in the model simulation. The apparent replication of these small oscillations makes

one wonder whether the two proxies might, in some ways, be rather precise thermometers.

Ultimately, we can have confidence in global temperature reconstructions only when they are based on comprehensive, multi-proxy data sets that span the Earth, and in which local and proxy-based noise is dampened. Marsicek and co-workers' synthesis of North American and European pollen records continues the march in this direction — these data were previously analysed within each continent to produce reconstructions^{6,7}, but they have now been carefully joined together and compared with a model. Further progress could be made through similar efforts to mine data obtained from around the world, much as has been done to refine the hockey-stick reconstructions¹.

Seasonal trends could also be better accounted for by focusing on proxies that have known seasonal biases, such as glaciers⁸, which often record summer temperatures, and permafrost ice wedges⁹, which reflect winter conditions. Moreover, reconstructions must be compared with more models, to examine how well they simulate feedbacks that could produce nonlinear responses to orbital forcing, such as the migration of the Arctic tree-line¹⁰ or changing atmospheric dust fluxes at low latitudes¹¹.

Finally, Marsicek and colleagues' study highlights two ways in which present climate change differs from that of the past. First, warming over the past century has occurred nearly everywhere and during all seasons, which contrasts with the more disparate spatial and seasonal trends during the Holocene. Second, if the Holocene was characterized by long-term temperature rises, as Marsicek *et al.* suggest, then the warming associated with human activities during the past century has probably already pushed temperatures beyond the range of temperatures that occurred during the Holocene. All the more reason, then, to improve reconstructions and models to better navigate that conundrum. ■

Jeremy D. Shakun is in the Department of Earth and Environmental Sciences, Boston College, Chestnut Hill, Massachusetts 02467, USA.

e-mail: jeremy.shakun@bc.edu

1. PAGES2k Consortium. *Sci. Data* **4**, 170088 (2017).
2. Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. *Science* **339**, 1198–1201 (2013).
3. Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L. & Brewer, S. *Nature* **554**, 92–96 (2018).
4. Laepple, T. & Lohmann, G. *Paleoceanography* **24**, PA4201 (2009).
5. Liu, Z. *et al.* *Proc. Natl Acad. Sci. USA* **111**, E3501–E3505 (2014).
6. Davis, B. A. S., Brewer, S., Stevenson, A. C. & Guiot, J. *Quat. Sci. Rev.* **22**, 1701–1716 (2003).
7. Viau, A. E., Gajewski, K., Sawada, M. C. & Fines, P. *J. Geophys. Res.* **111**, D09102 (2006).
8. Solomina, O. N. *et al.* *Quat. Sci. Rev.* **111**, 9–34 (2015).
9. Meyer, H. *et al.* *Nature Geosci.* **8**, 122–125 (2015).
10. MacDonald, G. M., Kremenetski, K. V. & Beilman, D. W. *Phil. Trans. R. Soc. B* **363**, 2283–2299 (2008).
11. Albani, S. *et al.* *Clim. Past* **11**, 869–903 (2015).

Enhancing the potential of enantioselective organocatalysis with light

Mattia Silvi¹ & Paolo Melchiorre^{2,3}

Organocatalysis—catalysis mediated by small chiral organic molecules—is a powerful technology for enantioselective synthesis, and has extensive applications in traditional ionic, two-electron-pair reactivity domains. Recently, organocatalysis has been successfully combined with photochemical reactivity to unlock previously inaccessible reaction pathways, thereby creating new synthetic opportunities. Here we describe the historical context, scientific reasoning and landmark discoveries that were essential in expanding the functions of organocatalysis to include one-electron-mediated chemistry and excited-state reactivity.

The preparation of chiral molecules with a well-defined, three-dimensional spatial arrangement is central to synthetic chemistry. Enantioselective organocatalysis offers powerful solutions to this challenging task¹. This strategy, which uses only small organic molecules as chiral catalysts, has greatly enriched the synthetic toolbox, complementing traditional metal-based and enzymatic approaches to asymmetric catalysis². Although sporadic examples of organic-catalyst-mediated processes were documented in the twentieth century^{3–6}, enantioselective organocatalysis gained prominence from 2000 onwards^{7,8}. A review⁹ published in 2008 suggested that the field of organocatalysis had developed so considerably in a relatively short period of time owing to the identification of a few generic mechanisms of substrate activation and stereochemical induction (detailed in Box 1), which provided effective tools for reaction invention. At that time⁹, organocatalysis was almost exclusively applied within traditional two-electron-pair reactivity domains, and reached high levels of efficiency, as evidenced by applications in the total synthesis of natural products^{10,11}. Because of this progress, the general perception within the chemistry community was that it would be difficult to further expand the synthetic potential of organocatalysis. However, this perception has been challenged by the merger of organocatalysis and photochemical reactivity¹², two powerful strategies for the activation of molecules that had previously remained largely distinct.

Here we outline the historical context and the scientific reasoning that motivated the combination of photocatalysis¹³ and organocatalysis. Instead of providing an exhaustive list of reactions, this Review critically describes developments since 2008, charting the essential ideas, serendipitous observations and landmark discoveries that were crucial in shifting organocatalysis beyond the established patterns of polar reactivity. A selection of pioneering studies will demonstrate how the merging of organocatalysis and light-mediated chemistry has had a profound influence on other fields of chemical research, such as radical chemistry¹⁴ and enantioselective photochemistry¹⁵ (see Box 2). In terms of stereoselectivity, impressive results have been achieved in many one-electron-mediated transformations, dispelling the long-held idea that the high reactivity of radicals limits their use in enantioselective catalysis¹⁶. Similarly, some organocatalytic tools have been used to enforce high levels of stereocontrol in photochemical processes, challenging the idea that photochemistry is too unselective to enable the efficient preparation of chiral molecules. This Review will also highlight the strong connections and developmental links between organocatalysis and the rapidly growing field of photoredox catalysis mediated by visible light¹⁷.

Merging organo- and photoredox catalysis

It is of interest to consider why organocatalysis was combined with photochemical reactivity, and what motivated the exploration beyond the established boundaries of two-electron-pair reactivity. As is often the case in science, progress was spurred by a specific goal that could not be achieved with the available technologies. Here, that goal was the intermolecular enantioselective α -alkylation of carbonyl substrates with alkyl halides (Fig. 1a) using an enamine-mediated catalytic process. It is important to understand why this simple transformation attracted great interest in the enantioselective-catalysis community¹⁸. The α -alkylation of carbonyl compounds is among the most important of the classical synthetic reactions¹⁹. Generally, the process requires the preformation of stoichiometric metal-enolate nucleophiles that undergo an S_N2 -type reaction with alkyl halides²⁰. The development of an enantioselective catalytic version, however, has proven difficult, with the few reported methodologies being limited in scope^{21,22}. Clearly, seeking to develop catalytic asymmetric methods that could directly functionalize unmodified carbonyl substrates was ambitious, and enamine-based chemistry was considered to be the most promising approach. This began with Gilbert Stork's fundamental studies²³ in the 1960s, which taught organic chemists that stoichiometric enamines could react with alkyl halides via S_N2 pathways. With the advent of enamine-mediated catalysis⁷, it was thought that implementing a direct stereoselective intermolecular α -alkylation of aldehydes would be not only feasible, but also straightforward. However, this synthetic target turned out to be much more difficult than expected²⁴. The main reason was the modest reactivity of alkyl halides, which complicates the ionic alkylation step while favouring side reactions, for example *N*-alkylation of the Lewis-basic amine catalysts and self-aldol condensation.

In 2008, David MacMillan's group recognized that the main hurdle to overcome was intrinsic to the ionic S_N2 pathway²⁵. Therefore, they used alkyl bromides not as electrophiles but as precursors for generating radicals. The underlying idea was to exploit the innate tendency of electron-deficient radicals to react rapidly with π -rich olefins, enabling the formation of carbon–carbon bonds that were otherwise difficult to make²⁶. A ruthenium-based polypyridyl photocatalyst **5** ($[\text{Ru}(\text{bpy})_3]^{2+}$, in which bpy is 2,2'-bipyridine) was used to generate open-shell species readily from α -bromo carbonyl compounds **2** (Fig. 1b). Before this point, photocatalyst **5** had a rich history as a single-electron transfer (SET) catalyst for inorganic applications²⁷, but had limited use in synthetic chemistry²⁸.

The reaction mechanism (Fig. 1c) is based on the integration of two independent catalytic cycles. On one side, the photoredox cycle proceeds

¹School of Chemistry, University of Bristol, Cantock's Close, Bristol BS8 1TS, UK. ²ICIQ, Institute of Chemical Research of Catalonia—the Barcelona Institute of Science and Technology, Avenida Països Catalans 16, 43007 Tarragona, Spain. ³ICREA, Catalan Institution for Research and Advanced Studies, Passeig Lluís Companys 23, 08010 Barcelona, Spain.

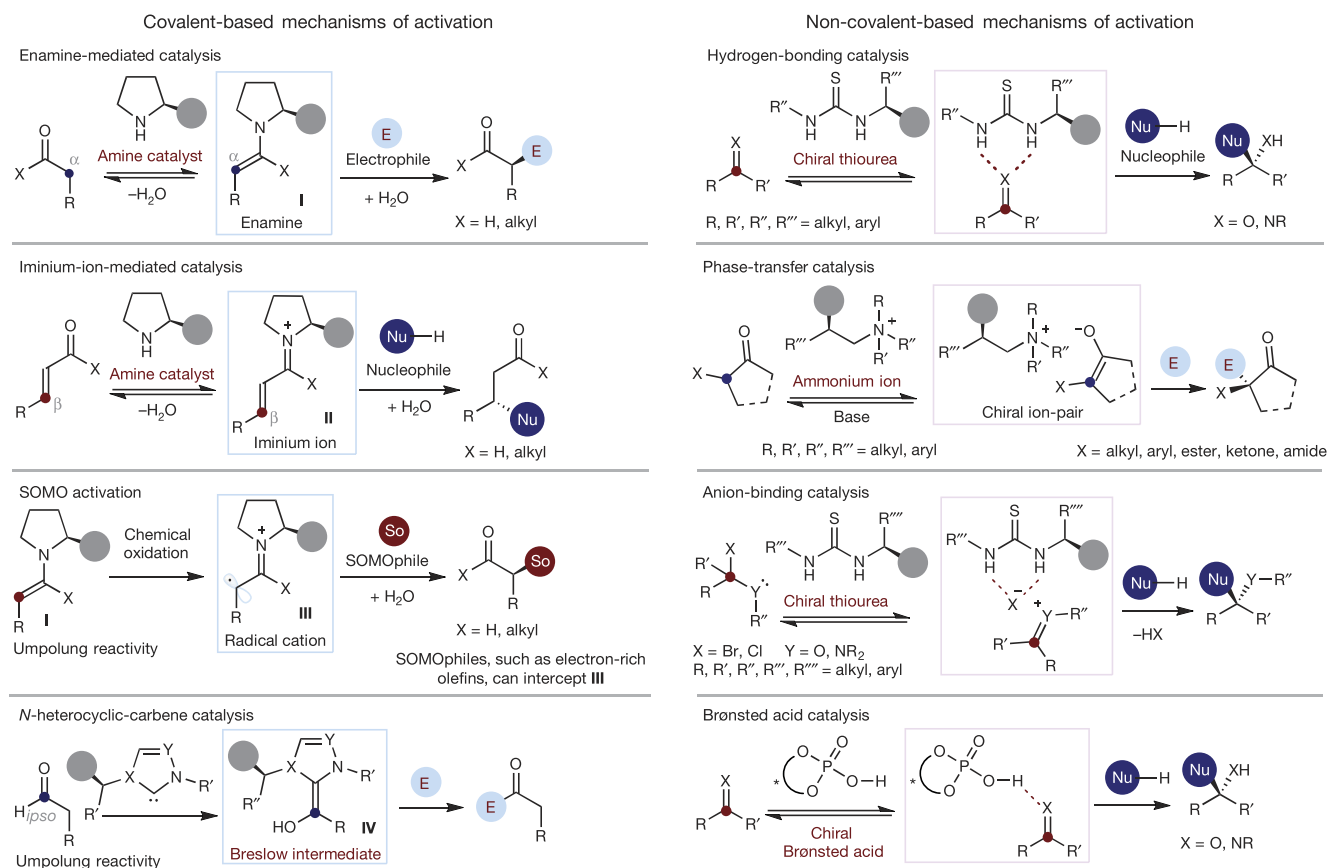
BOX 1

Generic mechanisms of organocatalytic reactivity

Organic catalysts can exert their functions by following two different substrate activation patterns (see Box 1 Figure):

Covalent-based modes of activation exploit the ability of an organic catalyst to covalently bind a substrate in a reversible fashion and form a reactive intermediate that can participate in many reaction types with consistently high enantioselectivity. Chiral primary and secondary amines belong to this class, activating carbonyl substrates via the formation of nucleophilic enamines **I** (refs 7, 87) (from enolizable aldehydes and ketones), electrophilic iminium ions **II** (refs 8, 88) (from unsaturated carbonyl compounds), and α -iminyl radical cation intermediates **III** (ref. 40) (upon single-electron oxidation of enamines by a chemical oxidant). *N*-heterocyclic carbene catalysts⁸⁹ offer an alternative activation mechanism for aldehydes, conferring an inverted (umpolung) reactivity to the normally electrophilic carbonyl carbon atom upon formation of Breslow intermediate **IV** (ref. 90), which acts as an acyl anion equivalent^{91,92}. These activation modes, which rely on strong, directional interactions, enable the stereoselective functionalization of unmodified carbonyl compounds at the *ipso*, α , and β positions.

Non-covalent approaches are based on the cooperation of several weak attractive interactions between the catalyst and a basic functional group on the substrates⁹³. Although the catalyst–substrate interactions are generally weaker and less directional than their covalent counterparts, they operate in concert to ensure a high level of transition state organization, resulting in a high degree of enantioselectivity. Hydrogen-bonding activation^{74,94}, phase-transfer catalysis^{6,70}, anion-binding activation⁹⁵, and Brønsted acid catalysis^{96–98} are all useful organocatalytic strategies for making chiral molecules⁹⁹.



Box 1 Figure | Organocatalytic mechanisms. Timelines: enamine-mediated catalysis: first applications 1970 (refs 3, 4), conceptualization 2000 (ref. 7), see ref. 87 for review. Iminium-mediated catalysis: first application 2000 (ref. 8), see ref. 88 for review. SOMO activation: first application 2007 (ref. 40). *N*-heterocyclic carbene catalysis: first application 1973 (ref. 91), conceptualization 2002 (ref. 92), see ref. 89 for review. Hydrogen-bonding catalysis: first application 1998 (ref. 94), see ref. 74 for review. Phase-transfer catalysis: first application 1984 (ref. 6), see ref. 70 for review. Anion-binding catalysis: first application 2007 (ref. 95), see ref. 99 for review. Brønsted acid catalysis: first application 2004 (refs 96, 97), see ref. 98 for review.

through the reductive cleavage of **2**, instigated by SET reduction from the Ru(I) intermediate **7**, to afford electrophilic radical **8**. Concurrently, in the organocatalytic pathway, enamine **1a** is generated by the condensation of organocatalyst **4** with aldehyde **1**. Then, the ground-state chiral enamine stereoselectively traps radical **8** to form the stereogenic centre within α -amino radical **9** with high fidelity. In the original study, it was proposed that this electron-rich intermediate **9** was finally oxidized by the excited

state of the Ru(II) photocatalyst **6**, a SET event that closes the photoredox cycle while affording iminium ion **10**. Hydrolysis of the latter species furnishes the α -alkylation product **3** while regenerating the catalyst **4**. Luminescence quenching studies revealed that the reducing [Ru(bpy)₃]⁺ species **7** was initially generated by the oxidation of a sacrificial amount of enamine **1a** by the excited *Ru(II) catalyst **6**. Later, mechanistic investigations established a radical chain manifold as the main reaction pathway²⁹

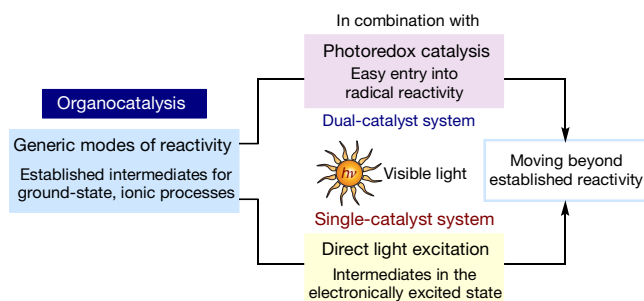
BOX 2

Light in organocatalysis

Two main strategies, the dual- and the single-catalyst approach, have been used to successfully combine organocatalysis and photochemical reactivity (see Box 2 Figure).

Dual-catalyst approach. In this strategy, the activity of a photoredox catalyst¹⁷ synergistically combines with the generic mechanisms of activation that define the ground-state reactivity of chiral organocatalytic intermediates. This approach involves the use of metal or organic photocatalysts that absorb light in the visible region; upon excitation, these photocatalysts either remove an electron from or donate an electron to simple organic substrates. This single-electron transfer mechanism facilitates access to radical species under mild conditions¹⁰⁰. The unique reactivity of such photocatalytically generated open-shell intermediates enables the expansion of the organocatalytic functions from a polar to a radical reactivity domain. The field of photoredox catalysis, which is a fast-moving area of modern synthetic chemistry¹⁷, has led to the development of many novel synthetic methodologies.

Single-catalyst approach. This strategy exploits the ability of organocatalytic intermediates to reach an excited state directly upon light absorption, and to participate in the activation of substrates, without external photocatalysts. At the same time, the chiral organocatalyst ensures effective stereochemical control. This approach demonstrates that the synthetic potential of organocatalytic intermediates is not limited to the ground-state domain, but can be expanded by exploiting their photochemical activity. By bringing an organocatalytic intermediate to an electronically excited state, light excitation unlocks reaction manifolds that are unavailable to conventional ground-state organocatalysis.



Box 2 Figure | Two main strategies for the combination of organocatalysis and photochemical reactivity.

(Fig. 1d). The photoredox catalyst therefore initiates a self-propagating radical process that is sustained by the ability of the α -amino radical **9** to regenerate radical **8** by directly reducing the organic bromide **2**. The same reaction can be achieved by replacement of the ruthenium photocatalyst with organic dyes³⁰ or different metal-based polypyridyl complexes³¹.

This study has had many far-reaching implications. Synthetically, by combining enamine-mediated catalysis with the action of a photoredox catalyst, mechanistically related enantioselective α -alkylation reactions have been developed (Fig. 1e), including trifluoromethylation³², benzylation³³, and cyanoalkylation³⁴ processes. The enantioselective α -alkylation of 1,3-dicarbonyl substrates to generate quaternary carbon stereocentres, which are synthetically useful yet difficult to form, has also been achieved³⁵. However, the main impact of this study was the demonstration that radical intermediates could be generated at ambient temperatures, simply by using a photocatalyst activated by visible light. This meant that the tools and the mechanisms for stereocontrol of enantioselective

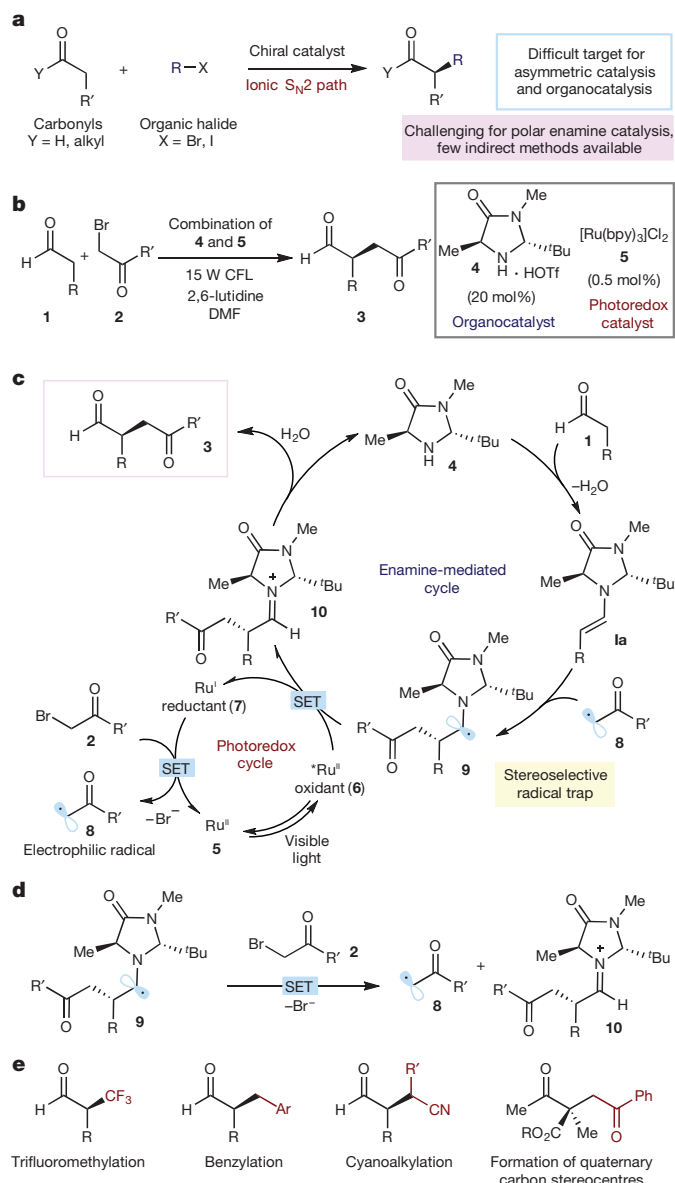


Figure 1 | Merging photoredox and enamine catalysis. **a**, The synthetic challenge of developing an intermolecular catalytic enantioselective α -alkylation of unmodified carbonyl substrates with alkyl halides via an ionic S_N2 path. **b**, Nicewicz and MacMillan's solution, involving the combination of enamine-mediated catalytic reactions and photoredox catalysis²⁵. **c**, The originally proposed closed catalytic cycle²⁵. **d**, The key propagation step of the radical chain mechanism²⁹. **e**, Further synthetic applications of this dual catalytic strategy for the direct stereocontrolled α -alkylation of aldehydes: trifluoromethylation³², benzylation³³, cyanoalkylation³⁴, and the formation of quaternary carbon stereocentres³⁵. CFL, compact fluorescence lamp.

organocatalysis, which require mild conditions for optimal efficiency, could be successfully applied within radical reactivity patterns. These studies, along with other investigations dealing with non-stereocontrolled transformations^{36,37}, also laid the foundations for the development of the field of photoredox catalysis¹⁷. At present, synthetic chemists are exploring the benefits of integrating the activity of photoredox catalysts with other catalytic systems, including metal-based catalysis³⁸ and chiral Lewis acid catalysis³⁹, although these aspects fall out of the scope of this Review.

Dual-catalyst systems—covalent organocatalysis

The use of redox-active photocatalysts in conjunction with well-established chiral organocatalytic intermediates other than enamines

has been successively explored. Two examples of such strategies are given below.

Merging SOMO activation and photoredox catalysis

The example of singly-occupied molecular orbital (SOMO) activation illustrates how the combination of established modes of organocatalytic reactivity with photoredox catalysis could lead to unconventional transformations. SOMO activation, first introduced in 2007 (ref. 40), exploits the SET oxidation of enamines **I** by a chemical oxidant, which generates an electrophilic α -iminyl radical cation **III** amenable to a range of open-shell reactions. Because **III** can be stereoselectively intercepted by electron-rich functionalized olefins (for example, allyl silanes), the subsequent α -alkylation products result from umpolung reactivity. The main drawback of this strategy is that it requires an excess of stoichiometric oxidant. This issue was solved using a light-activated catalyst to trigger the key SET oxidation of enamines to generate the intermediate **III** (ref. 41) (Fig. 2a). The milder radical-generation conditions offered the possibility of intercepting **III** with unactivated olefins, such as simple styrenes, in a stereocontrolled fashion⁴² (path i in Fig. 2a). By avoiding the use of organic halides, this approach further expanded the potential of the organocatalytic intermolecular α -alkylation technology. The chemistry required the combination of organocatalysis with both an iridium photoredox catalyst⁴³, which generated intermediate **III**, and a hydrogen atom transfer⁴⁴ thiol catalyst, which reduced the intermediate **V** resulting from the radical addition to the styrene.

The chemistry of α -iminyl radical cation **III**, generated under photoredox conditions, is not limited to radical addition manifolds. It can be expanded to realize unconventional and difficult-to-achieve transformations, such as the direct β -arylation of saturated carbonyl substrates⁴⁵ (path ii in Fig. 2a). The allylic C–H bonds in intermediate **III** are sufficiently weakened to enable proton abstraction by a suitable base, such as DABCO (1,4-diazabicyclo[2.2.2]octane), giving the β -enaminy radical intermediate **VI**. This species can undergo radical coupling with the long-lived radical anion **13**, generated by SET reduction of 1,4-dicyanobenzene **12** from an iridium(III) photocatalyst. This bond-forming event, which is governed by the persistent radical effect⁴⁶, forms a new carbon–carbon bond at the β -position of the original carbonyl. The strategy is synthetically appealing, given the lack of alternative methods for the direct β -functionalization of carbonyl substrates bearing saturated alkyl chains; however, only a single enantioselective example has been reported so far. Nonetheless, this study provided an initial demonstration that classical organocatalytic tools, such as the chiral amine catalyst **14**, could serve to control the stereochemical outcome of a radical coupling event, which is greatly complicated by its intrinsic high rate⁴⁷.

Overall, the studies detailed in Fig. 2a indicate that the native reactivity of an established organocatalytic intermediate (that is, enamines) can be switched from a closed-shell to an open-shell manifold with a light-activated photoredox catalyst. These studies also highlight the ability of traditional organic catalysts, generally used in enantioselective ionic processes, to control the geometry of the ensuing radical intermediates (such as **III** and **VI**) while creating a suitable chiral environment for stereocontrolled bond formation.

Merging iminium-ion and photoredox catalysis

Iminium-ion activation has found many applications in ionic domains, facilitating the conjugate additions of soft nucleophiles to the β -carbon atom of unsaturated carbonyl compounds. However, the development of a stereoselective trap of nucleophilic radicals has not been trivial. This is because the addition of radicals to a cationic iminium ion **II** creates a reactive α -iminyl radical cation **VII** (Fig. 2b), an unstable intermediate with a high tendency to undergo β -scission⁴⁸ and reform the more stable iminium ion **II**. Recently, a strategy was reported that enabled enantioselective radical conjugate additions to β -substituted cyclic enones **15**, forming quaternary carbon stereocentres with high fidelity⁴⁹. To

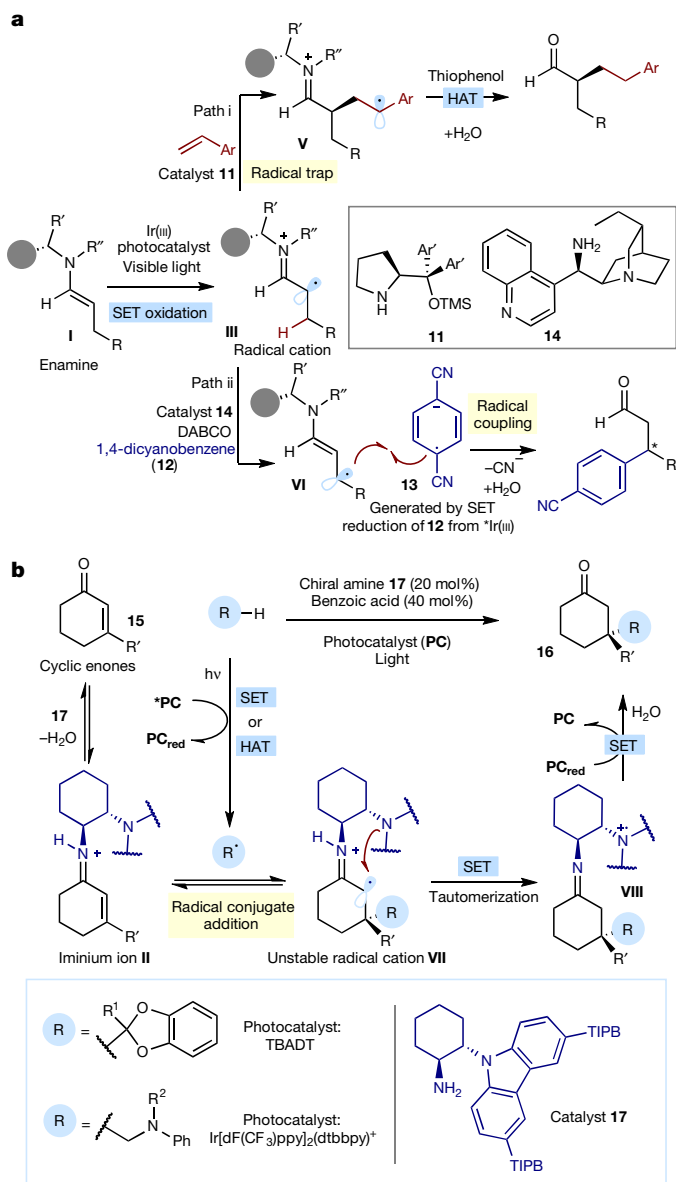


Figure 2 | Merging photoredox and covalent organocatalysis.

a, Irradiation of an iridium(III) photocatalyst generates an excited state that can remove an electron from the enamine (**I**) to afford the radical cation **III**. The chiral intermediate **III** can follow two different reaction manifolds: (i) it can be intercepted by styrenes to eventually afford α -homobenzylated aldehydes; (ii) it can be deprotonated to furnish the β -enaminy radical intermediate **VI**, which can then engage in a radical coupling with the radical anion **13**. In both cases, the stereodefining event is controlled by a chiral open-shell organocatalytic intermediate; Ar', 3,5-(CF₃)₂C₆H₃. **b**, To implement an iminium-ion-catalysed conjugate addition of radicals, the short-lived radical intermediate **VII** must be bypassed. This is achieved by intramolecular reduction from the electron-rich carbazole moiety within catalyst **17**. DABCO, 1,4-diazabicyclo[2.2.2]octane; HAT, hydrogen atom transfer; dtbbpy, 4,4'-bis(*tert*-butyl)-2,2'-bipyridine; dF(CF₃)ppy, 3,5-difluoro-2-[5-(trifluoromethyl)-2-pyridinyl]phenyl; LED, light-emitting diode; PC, photoredox catalyst; PC_{red}, reduced photocatalyst; TBABF₄, tetrabutylammonium tetrafluoroborate; TBADT, tetrabutylammonium decatungstate; TIPB, triisopropylbenzene; TMS, trimethylsilyl.

bypass the species **VII**, an electron-rich carbazole moiety was tethered at a strategic position on the chiral primary amine catalyst **17**, where it is poised to carry out a rapid intramolecular SET reduction of the unstable **VII**, preventing it from breaking down. A fast tautomerization of the transient enamine intermediate (not shown) leads to the more stable

imine **VIII**, thus avoiding a possible competitive back-electron transfer. Finally, the long-lived carbazole radical cation in **VIII**, arising from the intramolecular SET, undergoes single-electron reduction from the reduced photoredox catalyst (**PC_{red}** in Fig. 2b). This restores the neutral carbazole moiety while yielding the quaternary product **16**. Notably, a photocatalyst (**PC**) both creates the nucleophilic radical and promotes the final redox process, which was identified as the turnover-limiting step of the overall reaction⁵⁰.

This process provides a route to the generation of quaternary carbon stereocentres in an enantioselective manner, and exploits the tendency of radicals to connect structurally congested carbons because their reactivity is only marginally affected by steric factors⁵¹. However, radical-based catalytic enantioselective strategies had previously found limited application in the formation of quaternary carbon stereocentres³⁵, and were not mentioned in a recent comprehensive review of available methods⁵². It appears that organocatalysis, in combination with photoredox catalysis, may offer effective tools to better exploit the intrinsic merits of radical reactivity.

N-heterocyclic carbene catalysis can also be used in conjunction with photoredox catalysts⁵³. Although this approach has not yet been used to stereoselectively trap photochemically generated radicals, this target appears feasible.

Dual-catalyst systems—non-covalent organocatalysis

Non-covalent modes of organocatalytic reactivity have also been used with photoredox catalysis. So far there have been only a few reports, but these have offered solutions to important synthetic problems. The initial approaches used photochemical strategies to generate, *in situ*, reactive closed-shell species (for example, iminium ions⁵⁴ and singlet oxygen⁵⁵), which were successively intercepted by chiral organocatalytic intermediates. The first application of non-covalent organocatalysis in light-mediated radical chemistry provided a strategy for an asymmetric aza-pinacol cyclization⁵⁶ (Fig. 3a). The combination of the chiral phosphoric acid catalyst **20** and an iridium photoredox catalyst promoted the intramolecular reductive coupling between the ketone and hydrazone moieties within substrate **18** to furnish the *syn* 1,2-amino alcohol derivatives **19** with high enantioselectivity. The process is triggered by the formation of the ketyl radical **21**, which is generated by a concerted proton-coupled electron transfer (PCET) process⁵⁷ driven by the cooperation of the photoredox and organic catalysts. PCET involves the simultaneous transfer of a proton and an electron in a single elementary step to enable processes that would be precluded by sequential, discrete proton and electron transfer steps. In this specific case, the direct SET reduction of the aryl ketone in **18** by the iridium photocatalyst alone would not be feasible. The ketyl radical **21**, generated by PCET, is primed to cyclize into the hydrazone. Subsequent hydrogen atom transfer from a terminal reductant (Hantzsch dihydropyridine) to the generated hydrazyl radical leads to the final product **19**. The high level of enantiocontrol indicates that the neutral ketyl radical **21** could maintain a considerable association, via tight hydrogen-bonding interactions, with the coordinating phosphate anion of the chiral Brønsted acid **20** during the course of the stereo-defining cyclization. This study established the possibility of using concerted PCET to realize enantioselective radical processes by streamlining the preparation of radicals that are otherwise difficult to achieve. It also suggested the somewhat unexpected finding that the weak interactions inherent to non-covalent organocatalysis are appropriate for the selective binding of radical intermediates while channelling the resulting processes towards stereoselective manifolds.

Recently, Takashi Ooi expanded on this concept by using chiral *P*-spiro tetraaminophosphonium ion **25**, which could selectively bind the anion-radical **26** via ion-pairing interactions⁵⁸ (Fig. 3b). This approach required the concomitant action of an iridium photoredox catalyst to reduce *N*-sulfonyl aldimines **22** and oxidize *N,N*-arylaminoethanes **23**. The radical coupling of **26** and **27**, governed by the chiral ion pair, gave the amine product **24** in high enantioselectivity. This study further demonstrated that organocatalysis can provide effective approaches to

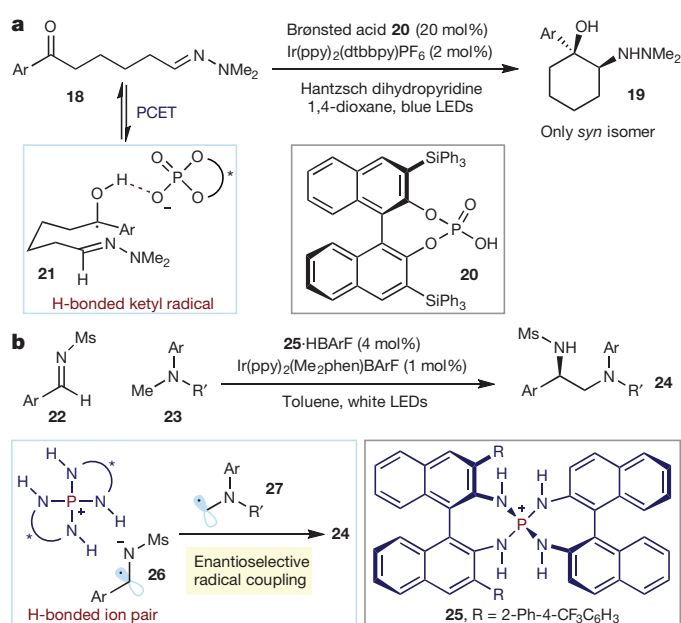


Figure 3 | Merging photoredox and non-covalent organocatalysis. **a**, The synergistic action of a chiral Brønsted acid and an iridium(III) photoredox catalyst facilitates both the formation of the ketyl radical **21**, by PCET, and the ensuing stereocontrolled aza-pinacol cyclization. **b**, The chiral ion-pair, formed between the cationic catalyst **25** and the photochemically generated radical anion **26**, governs an enantioselective radical coupling to afford products **24**. BARF, tetrakis[3,5-bis(trifluoromethyl)phenyl]borate; Ms, methanesulfonyl; ppy, 2-phenylpyridine.

address issues in enantioselective radical chemistry that were previously considered unattainable, such as the precise stereocontrol of radical coupling processes⁴⁷.

Organocatalysis in the excited state

The great potential of combining photoredox catalysis and organocatalysis lies mainly in the possibility of accessing open-shell species, the unique reactivity of which enables transformations not accessible through polar pathways. A different strategy has recently emerged, which offers possibilities to expand the field of organocatalysis. Researchers are exploring the potential of some chiral organocatalytic intermediates to reach an excited state directly upon the absorption of visible light, to enable new catalytic functions. The chemical reactivity of molecules differs fundamentally between the ground and electronically excited states⁵⁹. For example, a molecule in an excited state is both a better electron donor (that is, a better reductant) and a better electron acceptor (that is, a better oxidant) than it is in the ground state⁶⁰. This explains why, upon excitation, some organocatalytic intermediates can activate substrates via SET manifolds without the need for an external photocatalyst. At the same time, the chiral intermediate can provide effective stereochemical control over the ensuing bond-forming process. In this strategy, stereoreduction and photoactivation are combined in a single chiral organocatalyst.

Photochemistry of enamines

The reaction in Fig. 1b was also instrumental to the discovery that the synthetic potential of chiral enamines is not limited to the ground-state domain, and can be expanded by exploiting their photochemical activity. During investigations into the direct α -alkylation of aldehydes with electron-deficient alkyl bromides **28** using the organocatalyst **11** (Fig. 4a), a control experiment revealed that the reaction could be efficiently conducted in a stereoselective fashion under light illumination but without the need for any external photoredox catalyst⁶¹. Mechanistic studies revealed the ability of enamines **1b** to trigger the photochemical formation

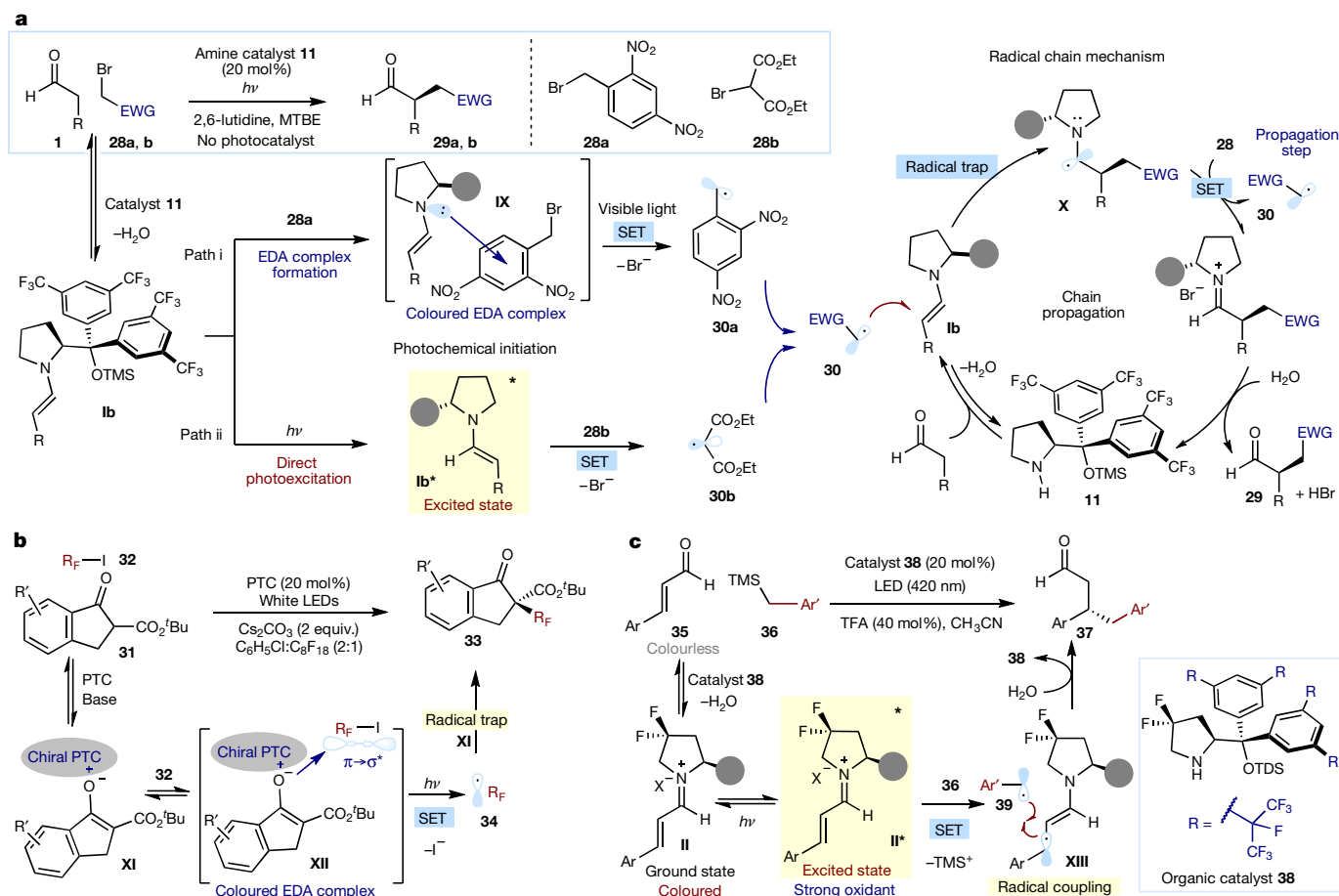


Figure 4 | Excited-state reactivity of chiral organocatalytic intermediates. **a**, Light-driven enantioselective α -alkylation of aldehydes: the photochemical activity of the enamines, either by EDA complex formation (path i) or direct photoexcitation (path ii), generates the electrophilic radicals **30** from electron-poor organic bromide **28**. The stereoselective radical trap, which is governed by the ground-state chiral enamine **1b**, triggers a chain propagation mechanism. **b**, Phase-transfer-catalysed enantioselective perfluoroalkylation of β -ketoesters, driven by the photoactivity of the enolate-based EDA complex **XII** (see ref. 71 for

PTC catalyst structure). **c**, Exploiting the direct photoexcitation of chiral iminium ions **II** to enable stereocontrolled β -alkylation of enals with non-nucleophilic alkyl silanes **36**; the chiral β -enaminy radical **XIII**, resulting from the SET reduction of the excited iminium ion **II***, governs the stereocontrolled radical coupling to afford products **37**. The filled grey circle represents the bulky chiral amine catalyst. EWG, electron-withdrawing group; PTC, phase-transfer catalyst; R_F , perfluoroalkyl fragment; TDS, tetrakis(dimethylsilyl); TFA, trifluoroacetic acid.

of radicals from alkyl bromides using two different photochemical mechanisms, depending on the substrate. The first mechanism⁶¹ relies on the formation of electron donor–acceptor (EDA) complexes that absorb visible light^{62,63}, which are generated in the ground state upon association of the electron-rich enamine **1b** with the electron-deficient dinitrobenzyl bromide **28a** (Fig. 4a, path i). Irradiation of the coloured EDA complex **IX** induces a SET event, enabling access to the radical intermediate **30a**. A second radical generation mechanism⁶⁴ (Fig. 4a, path ii) exploits the ability of the enamine **1b** to directly reach an electronically excited state (**Ib***) upon absorption of light in the near-UV region and then act as a potent single-electron reductant. SET reduction of bromomalonate **28b** induces the formation of radical **30b**. Mechanistic studies⁶⁵ established that both enamine-mediated photochemical alkylations proceeded through a self-propagating radical chain mechanism⁶⁶ (Fig. 4a, right), in analogy to the reactions in the presence of a photoredox catalyst (Fig. 1d).

These studies demonstrated that enamines, which behave as nucleophiles in the ground state, can become reductants upon light excitation and trigger the formation of radicals. At the same time, ground-state chiral enamines control the stereochemical course of the radical-trapping event. This strategy was expanded to develop mechanistically related enantioselective α -functionalization reactions, including phenacyl alkylation⁶¹, amination⁶⁷, and arylsulfonyl alkylation⁶⁸ of aldehydes and the alkylation of cyclic ketones⁶⁹.

Photochemistry of other organocatalytic intermediates

The discovery that the catalytic functions of enamines could be expanded by exploiting their excited-state reactivity⁶¹ motivated the quest for other chiral organocatalytic intermediates that could use similar photochemical mechanisms. Owing to their electronic similarities with enamines, enolates of type **XI**—generated *in situ* under phase-transfer catalysis conditions⁷⁰ (see Box 1) by deprotonation of cyclic β -ketoesters **31**—were suggested as suitable donors to facilitate the formation of photoactive ground-state EDA complexes⁷¹ (Fig. 4b). Perfluoroalkyl iodides (R_F I, **32**) served as electron-accepting substrates, leading to the formation of the coloured EDA complex **XII**. A single-electron transfer, promoted by visible light, triggered the formation of the perfluoroalkyl radical **34** ($R_F\cdot$) through the reductive cleavage of the C–I bond. The electrophilic nature of $R_F\cdot$ enabled it to be stereoselectively trapped by the chiral enolate **XI**, generated using a cinchonine-derived phase-transfer catalyst. This strategy provided access to enantio-enriched ketoester products **33** bearing a perfluoroalkyl-containing quaternary stereocentre.

Recently, it was also established that iminium ions can participate in photochemistry⁷² (Fig. 4c). Condensation of the chiral amine catalyst **38** with aromatic enals **35** converts an achromatic substrate into a coloured iminium ion **II**. Selective excitation with a violet light-emitting diode forms an electronically excited state (**II***), converting an electrophilic species into a strong oxidant⁷³ that can trigger the formation of radicals through SET oxidation of organic silanes **36**. The latter event furnishes

the β -enaminy radical intermediate **XIII** along with the radical **39**, which is generated upon irreversible fragmentation of the carbon–silicon bond. A stereocontrolled intermolecular coupling of the chiral β -enaminy radical **XIII** and **39** then forms the stereogenic centre in the β -functionalized aldehyde product **37**. The silane reagents **36** are non-nucleophilic substrates, which are recalcitrant to classical conjugate addition manifolds. Thus, in contrast to other examples of excited-state organocatalytic intermediates, the excitation of chiral iminium ions enables transformations that could not be realized by conventional catalytic asymmetric methodologies. A further difference is that the stereoselectivity is dictated by the chiral radical intermediate **XIII**, which governs the radical coupling event, and not by the ground-state iminium ion.

Non-covalent activation in asymmetric photochemistry

The photochemical organocatalytic strategies discussed so far all relied on the stereoselective interception of photogenerated radicals or radical ions in their ground states. But organocatalysis can also provide effective tools for catalytic stereocontrol in reactions of electronically excited intermediates. This is a difficult target because it requires the control of a photochemical process in a high-energy hypersurface, in which the action of a catalyst is greatly complicated by the absence of considerable activation barriers. Hydrogen-bonding catalysis⁷⁴, which relies on several weak interactions to activate the substrates, has provided effective solutions. Chiral ketones, appropriately adorned with hydrogen-bonding motifs⁷⁵, were used to catalyse light-triggered stereocontrolled cyclizations⁷⁶. The ketone-based organic catalysts effectively bind the substrate through a directional double-hydrogen-bond interaction, enabling the selective photoexcitation of a chiral catalyst–substrate complex. This ensures that the substrate resides in a suitable chiral environment when reaching an excited state. This strategy has been used successfully in both photo-induced redox processes and energy-transfer-induced photochemical reactions. In an example of the latter⁷⁷ (Fig. 5a), a visible-light-absorbing thioxanthone moiety was incorporated within the catalyst **42**. The lactam functionality of **42** was essential for binding the substrate **40** via a double-hydrogen-bond interaction. Meanwhile, the thioxanthone, upon excitation with light, activated the substrate via a proximity-driven Dexter energy transfer mechanism⁵⁹ and directed the [2 + 2] cyclization in the triplet energy hypersurface. The final product **41** was obtained with excellent enantioselectivity.

Other strategies for the enantioselective catalysis of photochemical processes^{15,78} have been successively developed. For example, it has been demonstrated that an intramolecular [2 + 2] photocycloaddition is promoted with high stereoselectivity by chiral thiourea catalysts⁷⁹, which are traditional ground-state hydrogen-bonding organocatalysts⁸⁰.

Photoexcitation of enzyme cofactors

Recently, a strategy has been reported that exploits the excited-state reactivity of common biological cofactors to enable enzymes to catalyse completely different processes than those for which they evolved. The natural reactivity of nicotinamide-dependent ketoreductases (KREDs) can be altered upon light excitation of the NADH/NADPH cofactor, which is bound to the enzyme active site⁸¹. KREDs have found extensive use in the preparation of chiral alcohols via the reduction of ketones⁸². This native polar reactivity arises from the ability of such enzymes to simultaneously bind, through non-covalent weak interactions, a carbonyl compound and the cofactor, and the tendency of NADH (or NADPH) to serve as a hydride source. Visible-light excitation, however, switches on a completely different reactivity in which the NAD(P)H becomes a strong reducing agent⁸³, enabling access to radical manifolds (Fig. 5b). This photochemical behaviour was used in the enantioselective dehalogenation of racemic α -bromo lactones **43**. Once the NAD(P)H and the substrate **43** are brought into close proximity in the active site of the enzyme, they can form a visible-light-absorbing EDA complex **XIV**, which triggers the formation of the prochiral radical intermediate **45** upon reductive cleavage of the substrate C–Br bond. The cofactor radical cation **46** drives the formation of the reduced chiral product **44**.

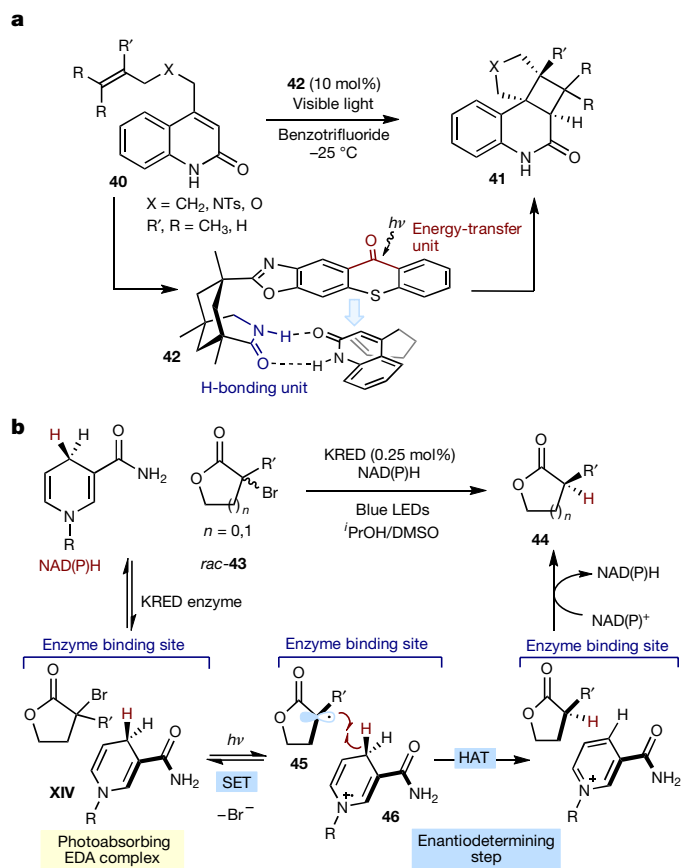


Figure 5 | Non-covalent interactions in enantioselective photochemistry. **a**, Hydrogen-bonding catalysis of enantioselective photochemical [2 + 2] cycloaddition via triplet energy-transfer mechanism. **b**, Photoexcitation of a NAD(P)H-dependent enzyme initiates a non-natural reactivity, enabling an enantioselective debromination of **43**. KRED, nicotinamide-dependent ketoreductase.

This brief detour into enzymatic catalysis⁸⁴ highlights how the power of photochemistry to unlock unconventional reactivity is influencing other established fields of catalytic enantioselective synthesis, including metal catalysis⁸⁵.

Conclusions and outlook

Over the last decade, combination with light has created exciting opportunities for expanding the scope of organocatalysis beyond conventional two-electron reactivity. Groundbreaking developments have taught synthetic chemists how to translate the generic mechanisms of activation, which govern the success of enantioselective polar organocatalysis, into the realm of excited-state reactivity and radical chemistry. The resulting light-driven methodologies are greatly expanding the way in which chemists consider the sustainable preparation of chiral molecules.

Major developments are probably still to come. This prediction is encouraged by the rapidly growing stream of innovation in photoredox catalysis, which is continuously offering new ways to generate radicals, and by the fact that the potential of excited-state organocatalytic reactivity is far from being fully revealed. Novel synthetic developments are expected to arise from the combination of photoredox catalysis and the activation mechanisms of ground-state organocatalysis. Considering the powerful photoredox methods available for the generation of radicals upon selective C–H activation of unactivated substrates (that is, PCET and hydrogen-atom-transfer mechanisms), the development of challenging enantioselective C(sp³)–C(sp³) coupling strategies is likely to become an ambitious target. Efforts will also be devoted to the use of continuous-flow photoreactors, which may enable the scale-up of photochemical organocatalytic asymmetric methods⁸⁶. Another central goal

for the continued expansion of organocatalysis will be to fully explore the unique modes of reactivity revealed by the excitation of organocatalytic intermediates. Along these lines, traditional photosensitizers could provide reliable support by facilitating—by means of energy-transfer mechanisms—the generation of excited-state chiral intermediates that cannot be accessed by the direct absorption of light. This approach will require a deep understanding of the photophysical properties of the organocatalytic intermediates. It is expected that the combination of conventional physical organic chemistry tools and photophysical investigations will play an increasingly relevant role. Another force for innovation may arise from the integration of the photochemical activity of chiral organocatalytic intermediates within metal-mediated catalytic cycles, which could enable unconventional mechanisms for stereocontrolled bond formation. Finally, we expect great advances in the development of photochemical radical cascade processes, in which the unique excited-state organocatalytic reactivities can be combined to provide powerful transformations for the one-step synthesis of complex chiral molecules¹⁰.

Received 8 August; accepted 24 November 2017.

- Dalko, P. I. (ed.) *Comprehensive Enantioselective Organocatalysis: Catalysts, Reactions, and Applications* (Wiley-VCH, 2013).
 - Ojima, I. (ed.) *Catalytic Asymmetric Synthesis* (John Wiley & Sons, 2010).
 - Eder, U., Sauer, G. & Wiechert, R. New type of asymmetric cyclization to optically active steroid CD partial structures. *Angew. Chem. Int. Ed. Engl.* **10**, 496–497 (1971).
 - Hajos, Z. G. & Parrish, D. R. Asymmetric synthesis of bicyclic intermediates of natural product chemistry. *J. Org. Chem.* **39**, 1615–1621 (1974).
 - Hiemstra, H. & Wynberg, H. Addition of aromatic thiols to conjugated cycloalkenones, catalyzed by chiral β -hydroxy amines. A mechanistic study of homogeneous catalytic asymmetric synthesis. *J. Am. Chem. Soc.* **103**, 417–430 (1981).
 - Dolling, U. H., Davis, P. & Grabowski, E. J. J. Efficient catalytic asymmetric alkylations. 1. Enantioselective synthesis of (+)-indacrinone via chiral phase-transfer catalysis. *J. Am. Chem. Soc.* **106**, 446–447 (1984).
 - List, B., Lerner, R. A. & Barbas, C. F., III. Proline-catalyzed direct asymmetric aldol reactions. *J. Am. Chem. Soc.* **122**, 2395–2396 (2000).
 - Ahrendt, K. A., Borths, C. J. & MacMillan, D. W. C. New strategies for organic synthesis: the first highly enantioselective organocatalytic Diels–Alder reaction. *J. Am. Chem. Soc.* **122**, 4243–4244 (2000).
- Refs 7 and 8 document the seminal studies on enamine- and iminium-ion-mediated catalysis, respectively, which established the field of modern organocatalysis.**
- MacMillan, D. W. C. The advent and development of organocatalysis. *Nature* **455**, 304–308 (2008).
- Thought-provoking discussion on the reasons behind the sudden growth in the field of modern organocatalysis.**
- Grondal, C., Jeanty, M. & Enders, D. Organocatalytic cascade reactions as a new tool in total synthesis. *Nat. Chem.* **2**, 167–178 (2010).
 - Jones, S. B., Simmons, B., Mastracchio, A. & MacMillan, D. W. C. Collective synthesis of natural products by means of organocascade catalysis. *Nature* **475**, 183–188 (2011).
 - Albini, A. & Fagnoni, M. (eds) *Handbook of Synthetic Photochemistry* (Wiley-VCH, 2010).
 - Schultz, D. M. & Yoon, T. P. Solar synthesis: prospects in visible light photocatalysis. *Science* **343**, 1239176 (2014).
 - Chatgililoglu, C. & Studer, A. (eds) *Encyclopedia of Radicals in Chemistry, Biology and Materials* (Wiley-VCH, 2014).
 - Brimioulle, R., Lenhart, D., Maturi, M. M. & Bach, T. Enantioselective catalysis of photochemical reactions. *Angew. Chem. Int. Ed.* **54**, 3872–3890 (2015).
- Insightful review of the general strategies and concepts underlying the implementation of photochemical homogenous catalytic enantioselective processes.**
- Sibi, M. P., Manyem, S. & Zimmerman, J. Enantioselective radical processes. *Chem. Rev.* **103**, 3263–3296 (2003).
 - Shaw, M. H., Twilton, J. & MacMillan, D. W. C. Photoredox catalysis in organic chemistry. *J. Org. Chem.* **81**, 6898–6926 (2016).
 - Melchiorre, P. Light in aminocatalysis: the asymmetric intermolecular α -alkylation of aldehydes. *Angew. Chem. Int. Ed.* **48**, 1360–1363 (2009).
 - Ireland, R. E. *Organic Synthesis* (Prentice-Hall, 1969).
 - Evans, D. A. in *Asymmetric Synthesis* Vol. 3, Part B (ed. Morrison, J. D.) Ch. 1, 1–110 (Academic, 1983).
 - Doyle, A. G. & Jacobsen, E. N. Enantioselective alkylations of tributyltin enolates catalyzed by Cr(salen)Cl: access to enantiomerically enriched all-carbon quaternary centers. *J. Am. Chem. Soc.* **127**, 62–63 (2005).
 - Dai, X., Strotman, N. A. & Fu, G. C. Catalytic asymmetric Hiyama cross-couplings of racemic α -bromo esters. *J. Am. Chem. Soc.* **130**, 3302–3303 (2008).

- Stork, G., Brizzolara, A., Landesman, H., Szmuszkovicz, J. & Terrell, R. The enamine alkylation and acylation of carbonyl compounds. *J. Am. Chem. Soc.* **85**, 207–222 (1963).
 - List, B. *et al.* The catalytic asymmetric α -benzylation of aldehydes. *Angew. Chem. Int. Ed.* **53**, 282–285 (2014).
 - Nicewicz, D. A. & MacMillan, D. W. C. Merging photoredox catalysis with organocatalysis: the direct asymmetric alkylation of aldehydes. *Science* **322**, 77–80 (2008).
- Besides providing a solution for the longstanding problem of the direct catalytic asymmetric α -alkylation of aldehydes, this seminal study demonstrated the great potential of combining photoredox and organocatalysis.**
- Giese, B. *Radicals in Organic Synthesis: Formation of Carbon–Carbon Bonds* (Pergamon, 1986).
 - Juris, A. *et al.* Ru(II) polypyridine complexes: photophysics, photochemistry, electrochemistry, and chemiluminescence. *Coord. Chem. Rev.* **84**, 85–277 (1988).
 - van Bergen, T. J., Hedstrand, D. M., Kruizinga, W. H. & Kellogg, R. M. Chemistry of dihydropyridine. 9. Hydride transfer from 1,4-dihydropyridine to sp^3 -hybridized carbon in sulfonium salts and activated halides. Studies with NAD(P)H models. *J. Org. Chem.* **44**, 4953–4962 (1979).
 - Cismesia, M. A. & Yoon, T. P. Characterizing chain processes in visible light photoredox catalysis. *Chem. Sci.* **6**, 5426–5434 (2015); erratum **6**, 6019 (2015).
- Early demonstration of the importance of applying classical experimental techniques, most relevant to photophysical investigations, for elucidating the mechanism of photoredox organocatalytic processes.**
- Neumann, M., Fuldner, S., König, B. & Zeitler, K. Metal-free, cooperative asymmetric organophotoredox catalysis with visible light. *Angew. Chem. Int. Ed.* **50**, 951–954 (2011).
 - Gualandi, A. *et al.* Organocatalytic enantioselective alkylation of aldehydes with [Fe(bpy)₃]Br₂ catalyst and visible light. *ACS Catal.* **5**, 5927–5931 (2015).
 - Nagib, D. A., Scott, M. E. & MacMillan, D. W. C. Enantioselective α -trifluoromethylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **131**, 10875–10877 (2009).
 - Shih, H.-W., Vander Wal, M. N., Grange, R. L. & MacMillan, D. W. C. Enantioselective α -benzylation of aldehydes via photoredox organocatalysis. *J. Am. Chem. Soc.* **132**, 13600–13603 (2010).
 - Welin, E. R., Warkentin, A. A., Conrad, J. C. & MacMillan, D. W. C. Enantioselective α -alkylation of aldehydes by photoredox organocatalysis: rapid access to pharmacophore fragments from β -cyanoaldehydes. *Angew. Chem. Int. Ed.* **54**, 9668–9672 (2015).
 - Zhu, Y., Zhang, L. & Luo, S. Asymmetric α -photoalkylation of β -ketocarbons by primary amine catalysis: facile access to acyclic all-carbon quaternary stereocenters. *J. Am. Chem. Soc.* **136**, 14642–14645 (2014).
 - Ischay, M. A., Anzovino, M. E., Du, J. & Yoon, T. P. Efficient visible light photocatalysis of [2 + 2] enone cycloadditions. *J. Am. Chem. Soc.* **130**, 12886–12887 (2008).
 - Narayanan, J. M. R., Tucker, J. W. & Stephenson, C. R. J. Electron-transfer photoredox catalysis: development of a tin-free reductive dehalogenation reaction. *J. Am. Chem. Soc.* **131**, 8756–8757 (2009).
 - Twilton, J., Le, C., Zhang, P., Shaw, M. H., Evans, R. W. & MacMillan, D. W. C. The merger of transition metal and photocatalysis. *Nat. Rev. Chem.* **1**, 0052 (2017).
 - Yoon, T. P. Photochemical stereocontrol using tandem photoredox–chiral Lewis acid catalysis. *Acc. Chem. Res.* **49**, 2307–2315 (2016).
 - Beeson, T. D., Mastracchio, A., Hong, J.-B., Ashton, K. & Macmillan, D. W. C. Enantioselective organocatalysis using SOMO activation. *Science* **316**, 582–585 (2007).
- Early use of organocatalysis in enantioselective radical chemistry, before the advent of photoredox catalysis.**
- Koike, T. & Akita, M. Photoinduced oxyamination of enamines and aldehydes with TEMPO catalyzed by [Ru(bpy)₃]²⁺. *Chem. Lett.* **38**, 166–167 (2009).
 - Capacci, A. G., Malinowski, J. T., McAlpine, N. J., Kuhne, J. & MacMillan, D. W. C. Direct, enantioselective α -alkylation of aldehydes using simple olefins. *Nat. Chem.* **9**, 1073–1077 (2017).
 - Flamigni, L., Barbieri, A., Sabatini, C., Ventura, B. & Barigelletti, F. in *Photochemistry and Photophysics of Coordination Compounds II* (eds Balzani, V. & Campagna, S.) 143–203 (Springer, 2007).
 - Mayer, J. M. Understanding hydrogen atom transfer: from bond strengths to Marcus theory. *Acc. Chem. Res.* **44**, 36–46 (2011).
 - Pirnot, M. T., Rankic, D. A., Martin, D. B. C. & MacMillan, D. W. C. Photoredox activation for the direct β -arylation of ketones and aldehydes. *Science* **339**, 1593–1596 (2013).
 - Studer, A. The persistent radical effect in organic synthesis. *Chem. Eur. J.* **7**, 1159–1164 (2001).
 - Curran, D. P., Porter, N. A. & Giese, B. (eds) *Stereochemistry of Radical Reactions* (VCH Verlag, 1996).
 - Jakobsen, H. J., Lawesson, S. O., Marshall, J. T. B., Schroll, G. & Williams, D. H. Mass spectrometry. XII. Mass spectra of enamines. *J. Chem. Soc. B* 940–946 (1966).
 - Murphy, J. J., Bastida, D., Paria, S., Fagnoni, M. & Melchiorre, P. Asymmetric catalytic formation of quaternary carbons by iminium ion trapping of radicals. *Nature* **532**, 218–222 (2016).
 - Bahamonde, A. *et al.* Studies on the enantioselective iminium ion trapping of radicals triggered by an electron-relay mechanism. *J. Am. Chem. Soc.* **139**, 4559–4567 (2017).

51. Fischer, H. & Radom, L. Factors controlling the addition of carbon-centered radicals to alkenes—an experimental and theoretical perspective. *Angew. Chem. Int. Ed.* **40**, 1340–1371 (2001).
52. Quasdorf, K. W. & Overman, L. E. Catalytic enantioselective synthesis of quaternary carbon stereocentres. *Nature* **516**, 181–191 (2014).
53. DiRocco, D. A. & Rovis, T. Catalytic asymmetric α -acylation of tertiary amines mediated by a dual catalysis mode: *N*-heterocyclic carbene and photoredox catalysis. *J. Am. Chem. Soc.* **134**, 8094–8097 (2012).
54. Bergonzini, G., Schindler, C. S., Wallentin, C.-J., Jacobsen, E. N. & Stephenson, C. R. J. Photoredox activation and anion binding catalysis in the dual catalytic enantioselective synthesis of β -amino esters. *Chem. Sci.* **5**, 112–116 (2014).
55. Lian, M., Li, Z., Cai, Y., Meng, Q. & Gao, Z. Enantioselective photooxygenation of β -keto esters by chiral phase-transfer catalysis using molecular oxygen. *Chem. Asian J.* **7**, 2019–2023 (2012).
56. Rono, L. J., Yayla, H. G., Wang, D. Y., Armstrong, M. F. & Knowles, R. R. Enantioselective photoredox catalysis enabled by proton-coupled electron transfer: development of an asymmetric aza-pinacol cyclization. *J. Am. Chem. Soc.* **135**, 17735–17738 (2013).
- Seminal example demonstrating the possibility of exploiting proton-coupled electron transfer in enantioselective organocatalysis.**
57. Miller, D. C., Tarantino, K. T. & Knowles, R. R. Proton-coupled electron transfer in organic synthesis: fundamentals, applications, and opportunities. *Top. Curr. Chem.* **374**, 30 (2016).
58. Uraguchi, D., Kinoshita, N., Kizu, T. & Ooi, T. Synergistic catalysis of ionic Brønsted acid and photosensitizer for a redox neutral asymmetric α -coupling of *N*-arylaminoethanes with aldimines. *J. Am. Chem. Soc.* **137**, 13768–13771 (2015).
59. Turro, N. J., Ramamurthy, V. & Scaiano, J. C. *Modern Molecular Photochemistry of Organic Molecules* (University Science Books, 2010).
60. Balzani, V., Ceroni, P. & Juris, A. *Photochemistry and Photophysics* (Wiley-VCH, 2014).
61. Arceo, E., Jurberg, I. D., Alvarez-Fernández, A. & Melchiorre, P. Photochemical activity of a key donor–acceptor complex can drive stereoselective catalytic α -alkylation of aldehydes. *Nat. Chem.* **5**, 750–756 (2013).
- First demonstration that enamines—key intermediates in ground-state organocatalysis—can use photochemical mechanisms to activate substrates.**
62. Mulliken, R. S. Molecular compounds and their spectra. II. *J. Am. Chem. Soc.* **74**, 811–824 (1952).
63. Rathore, R. & Kochi, J. K. Donor/acceptor organizations and the electron-transfer paradigm for organic reactivity. *Adv. Phys. Org. Chem.* **35**, 193–318 (2000).
64. Silvi, M., Arceo, E., Jurberg, I. D., Cassani, C. & Melchiorre, P. Enantioselective organocatalytic alkylation of aldehydes and enals driven by the direct photoexcitation of enamines. *J. Am. Chem. Soc.* **137**, 6120–6123 (2015).
65. Bahamonde, A. & Melchiorre, P. Mechanism of the stereoselective α -alkylation of aldehydes driven by the photochemical activity of enamines. *J. Am. Chem. Soc.* **138**, 8019–8030 (2016).
66. Studer, A. & Curran, D. P. Catalysis of radical reactions: a radical chemistry perspective. *Angew. Chem. Int. Ed.* **55**, 58–102 (2016).
67. Cecere, G., König, C. M., Alleva, J. L. & MacMillan, D. W. C. Enantioselective direct α -amination of aldehydes via a photoredox mechanism: a strategy for asymmetric amine fragment coupling. *J. Am. Chem. Soc.* **135**, 11521–11524 (2013).
68. Filippini, G., Silvi, M. & Melchiorre, P. Enantioselective formal α -methylation and α -benzylation of aldehydes by means of photo-organocatalysis. *Angew. Chem. Int. Ed.* **56**, 4447–4451 (2017).
69. Arceo, E., Bahamonde, A., Bergonzini, G. & Melchiorre, P. Enantioselective direct α -alkylation of cyclic ketones by means of photo-organocatalysis. *Chem. Sci.* **5**, 2438–2442 (2014).
70. Shirakawa, S. & Maruoka, K. Recent developments in asymmetric phase-transfer reactions. *Angew. Chem. Int. Ed.* **52**, 4312–4348 (2013).
71. Woźniak, Ł., Murphy, J. J. & Melchiorre, P. Photo-organocatalytic enantioselective perfluoroalkylation of β -ketoesters. *J. Am. Chem. Soc.* **137**, 5678–5681 (2015).
72. Silvi, M., Verrier, C., Rey, Y. P., Buzzetti, L. & Melchiorre, P. Visible-light excitation of iminium ions enables the enantioselective catalytic β -alkylation of enals. *Nat. Chem.* **9**, 868–873 (2017).
73. Mariano, P. S. Electron-transfer mechanisms in photochemical transformations of iminium salts. *Acc. Chem. Res.* **16**, 130–137 (1983).
74. Taylor, M. S. & Jacobsen, E. N. Asymmetric catalysis by chiral hydrogen-bond donors. *Angew. Chem. Int. Ed.* **45**, 1520–1543 (2006).
75. Bach, T., Bergmann, H., Grosch, B. & Harms, K. Highly enantioselective intra- and intermolecular [2 + 2] photocycloaddition reactions of 2-quinolones mediated by a chiral lactam host: host–guest interactions, product configuration, and the origin of the stereoselectivity in solution. *J. Am. Chem. Soc.* **124**, 7982–7990 (2002).
76. Bauer, A., Westkämper, F., Grimme, S. & Bach, T. Catalytic enantioselective reactions driven by photoinduced electron transfer. *Nature* **436**, 1139–1140 (2005).
- Seminal example of enantioselective organocatalysis of photochemical reactions in the excited state.**
77. Alonso, R. & Bach, T. A chiral thioxanthone as an organocatalyst for enantioselective [2 + 2] photocycloaddition reactions induced by visible light. *Angew. Chem. Int. Ed.* **53**, 4368–4371 (2014).
78. Brimioulle, R. & Bach, T. Enantioselective Lewis acid catalysis of intramolecular enone [2 + 2] photocycloaddition reactions. *Science* **342**, 840–843 (2013).
79. Vallavoju, N., Selvakumar, S., Jockusch, S., Sibi, M. P. & Sivaguru, J. Enantioselective organo-photocatalysis mediated by atropisomeric thiourea derivatives. *Angew. Chem. Int. Ed.* **53**, 5604–5608 (2014).
80. Madarász, Á. *et al.* Thiourea derivatives as Brønsted acid organocatalysts. *ACS Catal.* **6**, 4379–4387 (2016).
81. Emmanuel, M. A., Greenberg, N. R., Oblinsky, D. G. & Hyster, T. K. Accessing non-natural reactivity by irradiating nicotinamide-dependent enzymes with light. *Nature* **540**, 414–417 (2016).
- Landmark demonstration that light excitation of cofactors can alter the natural reactivity of enzymes.**
82. Huisman, G. W., Liang, J. & Krebber, A. Practical chiral alcohol manufacture using ketoreductases. *Curr. Opin. Chem. Biol.* **14**, 122–129 (2010).
83. Fukuzumi, S., Hironaka, K. & Tanaka, T. Photoreduction of alkyl halides by an NADH model compound. An electron transfer chain mechanism. *J. Am. Chem. Soc.* **105**, 4722–4727 (1983).
84. Bornscheuer, U. T. *et al.* Engineering the third wave of biocatalysis. *Nature* **485**, 185–194 (2012).
85. Huo, H. *et al.* Asymmetric photoredox transition-metal catalysis activated by visible light. *Nature* **515**, 100–103 (2014).
86. Cambié, D., Bottecchia, C., Straathof, N. J. W., Hessel, V. & Noël, T. Applications of continuous-flow photochemistry in organic synthesis, material science, and water treatment. *Chem. Rev.* **116**, 10276–10341 (2016).
87. Mukherjee, S., Yang, J. W., Hoffmann, S. & List, B. Asymmetric enamine catalysis. *Chem. Rev.* **107**, 5471–5569 (2007).
88. Lelais, G. & MacMillan, D. W. C. Modern strategies in organic catalysis: the advent and development of iminium activation. *Aldrichimica Acta* **39**, 79–87 (2006).
89. Enders, D., Niemeier, O. & Henseler, A. Organocatalysis by *N*-heterocyclic carbenes. *Chem. Rev.* **107**, 5606–5655 (2007).
90. Breslow, R. On the mechanism of thiamine action. IV. Evidence from studies on model systems. *J. Am. Chem. Soc.* **80**, 3719–3726 (1958).
91. Sheehan, J. & Hara, T. Asymmetric thiazolium salt catalysis of the benzoin condensation. *J. Org. Chem.* **39**, 1196–1199 (1974).
92. Enders, D. & Kalfass, U. An efficient nucleophilic carbene catalyst for the asymmetric benzoin condensation. *Angew. Chem. Int. Ed.* **41**, 1743–1745 (2002).
93. Knowles, R. R. & Jacobsen, E. N. Attractive noncovalent interactions in asymmetric catalysis: links between enzymes and small molecule catalysts. *Proc. Natl. Acad. Sci. USA* **107**, 20678–20685 (2010).
94. Sigman, M. & Jacobsen, E. N. Schiff base catalysts for the asymmetric Strecker reaction identified and optimized from parallel synthetic libraries. *J. Am. Chem. Soc.* **120**, 4901–4902 (1998).
95. Reisman, S. E., Doyle, A. G. & Jacobsen, E. N. Enantioselective thiourea-catalyzed additions to oxocarbenium ions. *J. Am. Chem. Soc.* **130**, 7198–7199 (2008).
96. Akiyama, T., Itoh, J., Yokota, K. & Fuchibe, K. Enantioselective Mannich-type reaction catalyzed by a chiral Brønsted acid. *Angew. Chem. Int. Ed.* **43**, 1566–1568 (2004).
97. Uraguchi, D. & Terada, M. Chiral Brønsted acid-catalyzed direct Mannich reactions via electrophilic activation. *J. Am. Chem. Soc.* **126**, 5356–5357 (2004).
98. Parmar, D., Sugiono, E., Raja, S. & Rueping, M. Complete field guide to asymmetric BINOL-phosphate derived Brønsted acid and metal catalysis: history and classification by mode of activation; Brønsted acidity, hydrogen bonding, ion pairing, and metal phosphates. *Chem. Rev.* **114**, 9047–9153 (2014).
99. Brak, K. & Jacobsen, E. N. Asymmetric ion-pairing catalysis. *Angew. Chem. Int. Ed.* **52**, 534–561 (2013).
100. Staveness, D., Bosque, I. & Stephenson, C. R. J. Free radical chemistry enabled by visible light-induced electron transfer. *Acc. Chem. Res.* **49**, 2295–2306 (2016).

Acknowledgements P.M. thanks the Generalitat de Catalunya (CERCA Program), Agencia Estatal de Investigación (AEI) (CTQ2016-75520-P), and the European Research Council (ERC 681840-CATA-LUX) for financial support. M.S. thanks the EU for a Horizon 2020 Marie Skłodowska-Curie Fellowship (grant 744242).

Author Contributions P.M. outlined the content of the Review and defined its scope. M.S. and P.M. worked together to prepare and edit the manuscript, figures and references.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to P.M. (pmelchiorre@iciq.es).

The axolotl genome and the evolution of key tissue formation regulators

Sergej Nowoshilow^{1,2,3,†*}, Siegfried Schloissnig^{4*}, Ji-Feng Fei^{5*}, Andreas Dahl^{3,6}, Andy W. C. Pang⁷, Martin Pippel⁴, Sylke Winkler¹, Alex R. Hastie⁷, George Young⁸, Juliana G. Roscito^{1,9,10}, Francisco Falcon¹¹, Dunja Knapp³, Sean Powell⁴, Alfredo Cruz¹¹, Han Cao⁷, Bianca Habermann¹², Michael Hiller^{1,9,10}, Elly M. Tanaka^{1,2,3,†} & Eugene W. Myers^{1,10}

Salamanders serve as important tetrapod models for developmental, regeneration and evolutionary studies. An extensive molecular toolkit makes the Mexican axolotl (*Ambystoma mexicanum*) a key representative salamander for molecular investigations. Here we report the sequencing and assembly of the 32-gigabase-pair axolotl genome using an approach that combined long-read sequencing, optical mapping and development of a new genome assembler (MARVEL). We observed a size expansion of introns and intergenic regions, largely attributable to multiplication of long terminal repeat retroelements. We provide evidence that intron size in developmental genes is under constraint and that species-restricted genes may contribute to limb regeneration. The axolotl genome assembly does not contain the essential developmental gene *Pax3*. However, mutation of the axolotl *Pax3* paralogue *Pax7* resulted in an axolotl phenotype that was similar to those seen in *Pax3*^{-/-} and *Pax7*^{-/-} mutant mice. The axolotl genome provides a rich biological resource for developmental and evolutionary studies.

Salamanders boast an illustrious history in biological research as the animal in which the Spemann organizer¹ and Sperry's chemoaffinity theory of axonal guidance² were discovered. Since 1768, when Spallanzani discovered tail and limb regeneration, researchers have probed this animal's remarkable regenerative capabilities with increasing molecular resolution. *A. mexicanum* (Fig. 1a) was first collected by von Humboldt, and has been cultivated in the laboratory since 1864 as a model for investigating phenomena such as nuclear reprogramming, the embryology of germ-cell induction, retinal neuron processing and regeneration³. Owing to the ease with which *A. mexicanum* can be bred in the laboratory, a sophisticated molecular toolkit has been developed for this species, including germline transgenesis and CRISPR-mediated gene mutation as well as viral and other transfection methods. These tools have enabled discoveries such as the identification of the source cells of regeneration and molecular pathways that control regeneration^{4,5}. A full exploitation of the axolotl model, including understanding regeneration and why it is limited in other tetrapods, requires analysis of its genome regulation and evolution. However, efforts towards comprehensive assembly of salamander genomes have been challenging owing to their large genome sizes (14–120 Gb) and the large number of repetitive regions they contain; the 32-Gb axolotl genome is ten times the size of the human genome. Here we report the sequencing, assembly and analysis of the axolotl genome.

A long-read assembler for large genomes

Our aim was to generate a genome sequence assembly for the D/D axolotl strain (Fig. 1a), which is commonly used in laboratory regeneration studies owing to its compatibility with live imaging. Taking into consideration the expected challenge of assembling the complex 32-Gb genome⁶, we sequenced 110 million long reads (32× coverage, N50 read length

14.2 kb) using Pacific Biosciences (PacBio) instruments (Supplementary Information section 1) to avoid the read sampling bias that is often found when using other technologies and to span repeat-rich genomic regions that cause breaks in short-read assemblies (Fig. 1b, c).

We developed an assembly algorithm (MARVEL) that integrates a two-phase read-correction procedure that keeps long PacBio reads intact for assembly (Supplementary Information section 2). MARVEL produced a contig assembly with an N50 of 218 kb. Next, we used 7× Illumina-based sequencing to correct sequence errors in 1% of the contig bases (Fig. 1b), which yielded a sequence accuracy of more than 99.2%. On the basis of the Illumina data, we estimated a heterozygosity of 0.47% (Supplementary Information section 2.2).

To provide a scaffold for the contig assembly, we generated *de novo* optical maps using the Bionano Saphyr system (Supplementary Information section 2.3). The Bionano map resolved contig chimaeras, which were found in 1.7% of contigs, slightly reducing N50 contig length to 216 kb (Fig. 1d). The final hybrid assembly yielded an N50 scaffold length of 3 Mb. Compared to the short-read assembly of the 20-Gb spruce genome⁷ or the 22-Gb loblolly pine genome⁸, which involved 12× long-read coverage, the axolotl assembly showed 56- and 29-fold improvements in contiguity, respectively (Table 1).

To assess the completeness of the assembly (Supplementary Information section 4.1), we first determined the number of aligning non-exonic ultraconserved elements⁹ (UCEs). We found that 194 (98.5%) of 197 non-exonic UCEs that are conserved across vertebrates align to the axolotl assembly. By comparison, 189 and 192 UCEs align to the Tibetan frog and *Xenopus* genomes, respectively, and 195 UCEs align to the coelacanth genome, indicating that the completeness of the axolotl genome assembly is comparable to or better than the two other amphibian genomes, which are smaller than 2.3 Gb¹⁰.

¹Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany. ²Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria. ³DFG Research Center for Regenerative Therapies, Technische Universität Dresden, Dresden, Germany. ⁴Heidelberg Institute for Theoretical Studies, Heidelberg, Germany. ⁵Institute for Brain Research and Rehabilitation, South China Normal University, Guangzhou, China. ⁶Deep Sequencing Group, Biotechnology Center (Biotec) Technische Universität Dresden, Dresden, Germany. ⁷Bionano Genomics, San Diego, California, USA. ⁸The Francis Crick Institute, London, UK. ⁹Max Planck Institute for the Physics of Complex Systems, Dresden, Germany. ¹⁰Center for Systems Biology, Dresden, Germany. ¹¹Molecular and Developmental Complexity Group, Unidad de Genómica Avanzada, Languebio-Cinvestav, Irapuato, Mexico. ¹²IBDM – Institut de Biologie du Développement de Marseille, CNRS & Aix-Marseille Université, Marseille, France. [†]Present address: Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Campus Vienna Biocenter 1, 1030 Vienna, Austria.

*These authors contributed equally to this work.

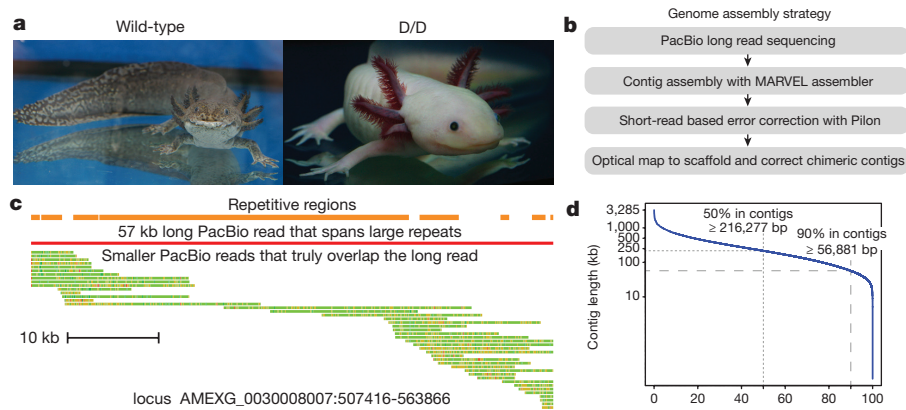


Figure 1 | Contiguity and completeness of the axolotl genome assembly. **a**, A wild-type *A. mexicanum* and the sequenced D/D *A. mexicanum* strain. **b**, The assembly strategy combines long-read sequencing, a novel assembler (MARVEL), error correction and scaffolding. **c**, A 57,385-bp PacBio read (red line) spans a large repetitive region (repeats are shown in orange; the

longest repeat is 34 kb) and, together with the other long reads shown below the long PacBio read, allows assembly of the locus (green-to-red colouring indicates alignment quality; repeat-induced alignments of reads belonging to other loci have been removed). **d**, $N(x)$ plot shows the percentage of the genome (x axis) that consists of contigs of at least x kb (y axis).

To further assess the completeness of the assembly, we generated a comprehensive gene catalogue by sequencing mRNA from 22 tissues (Supplementary Information section 3). Tissue-specific transcriptome assemblies and a composite assembly of all 1.5 billion transcript reads resulted in 180,649 transcript contigs (Supplementary Table 6) that contained 99% of the conserved core eukaryotic genes¹¹ and achieved the highest BUSCO score (<http://busco.ezlab.org/>) of an axolotl transcriptome reported to date (Supplementary Information section 3.4). More than 85% of the transcripts aligned to the genome along at least 95% of their length (Supplementary Information section 3.5), confirming the high completeness of the assembly. Furthermore, 71% of transcript contigs in which more than 95% of the sequence aligned with the genome were located on single scaffolds, demonstrating the high contiguity of the assembly. Using this comprehensive transcript set, we annotated a total of 23,251 protein-coding genes in the axolotl genome, a similar number to those found in other vertebrate genomes (Supplementary Information section 4.2).

Expansion of long terminal repeat retroelements

Given the similar number of genes in the *A. mexicanum* genome in comparison to other smaller vertebrate genomes, we analysed repetitive

sequences (Supplementary Information sections 4.2.2, 4.2.3). Repetitive sequences made up 65.6% of the contig assembly, representing a total of 18.6 Gb. Distinct long terminal repeat (LTR) retroelement classes and endogenous retroviruses made up the largest portion of the repetitive sequences (Fig. 2a, b, Supplementary Table 13) and included elements of more than 10 kb in length (Fig. 2c, Extended Data Fig. 1). Such long elements pose challenges for assembly, and indeed 97% of contigs ended in LTR elements. The number of substitutions to the repeat consensus, which is an estimate of the relative age of the LTR retroelement, indicates that the axolotl genome has undergone a long period of transposon activity followed by a recent and apparently continuing burst of expansion (Fig. 2d). This profile is consistent with previous small-scale characterizations of other salamander genomes¹².

The presence of many repeated elements contributes to a median intron size (22,759 bp) 13, 16 and 25 times that observed in human (1,750 bp), mouse (1,469 bp) and frog (906 bp), respectively (Fig. 3a, Supplementary Information section 4.3), a trend that was previously observed in five genes obtained from selective bacterial artificial chromosome sequencing of the axolotl genome¹³. Figure 3b shows a typical gene organization in axolotl compared to its human orthologue. Consistent with intron expansion, a distance comparison of pairs of highly conserved non-exonic elements shows that intergenic regions in the axolotl genome are 12 to 17 times larger than those in human, mouse and frog (Supplementary Information section 4.4).

HoxA cluster and intron size constraints

To examine gene cluster organization within this large genomic context, we focused on the HoxA locus, which has an important role in proximal-to-distal limb development and is reactivated during limb regeneration^{14,15}. The entire HoxA locus is contained on a single contig (Fig. 3c), and the conserved neighbouring gene *Evx1* is contained on the same 3.34-Mb scaffold. Compared to the orthologous human and frog clusters, the *A. mexicanum* HoxA cluster has a substantially increased repeat content and is 3.5 times larger, mostly owing to a 170-kb expansion between *HoxA3* and *HoxA4* (Fig. 3c). Notably, highly conserved non-exonic elements that putatively overlap *cis*-regulatory elements are not interspersed in this 170-kb region, but remain in proximity to *HoxA3* and *HoxA4*. The axolotl has a typical HoxA gene structure, with two coding exons separated by an intron. Notably, in contrast to the overall expansion of intron sizes, the intron sizes in the axolotl HoxA locus are very similar to those in other vertebrates, with the exception of *AmHoxA3*, which is also the longest of the HoxA genes in other tetrapods (Supplementary Table 17). Selected HoxC and HoxD genes examined in the red spotted newt exhibited similar properties¹⁶.

Table 1 | Comparison of assembly contiguity statistics in axolotl, spruce and pine genomes

	Axolotl (<i>A. mexicanum</i>)	White spruce (<i>Picea glauca</i>)	Loblolly pine (<i>Pinus taeda</i>)
Assembly size (Gb)	32.4 (28.4 in contigs)	24.6	20.6
Genome size (Gb)	32	20	22
Chromosomes	14	12	12
Sequencing technology	PacBio; Optical map	Illumina; cDNAs	Illumina; PacBio; Fosmid DiTag
Coverage	32×	65×	68× Illumina; 12× PacBio
Assembler	MARVEL	ABYSS	MaSuRCA
Contig N50 (bp)	216,277	6,644	25,361
Number of contigs	217,461	5,252,090	2,445,689
Scaffold N50 (bp)	3,052,786	54,661	107,036
Number of scaffolds	125,724	3,033,322	1,496,869

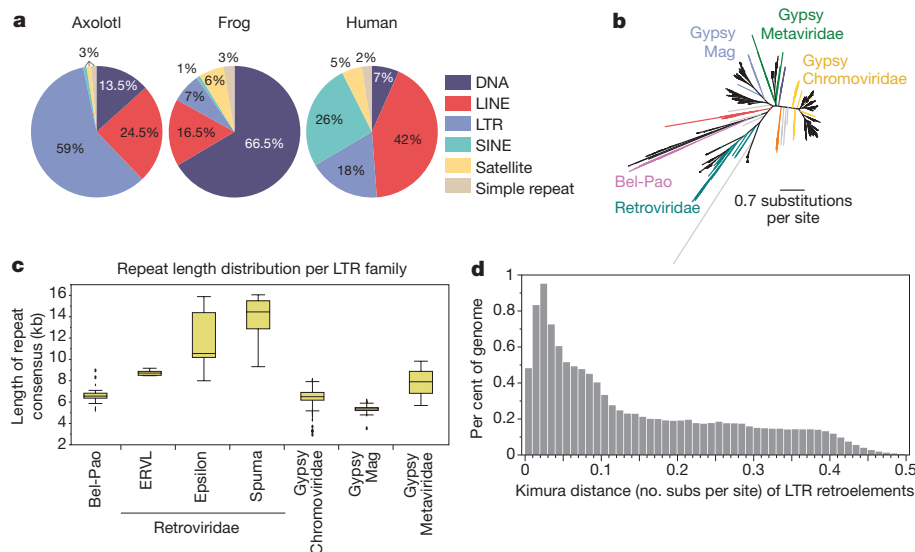


Figure 2 | The axolotl genome contains an expansion of LTR retroelements. **a**, Pie charts of major repeat classes (LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements) show an abundance of LTR elements. **b**, Phylogenetic tree of axolotl LTR-element clusters (black) and all LTR elements from GyDB2.0³⁷. Annotated clusters are indicated by colour, non-annotated clusters are in grey. Note that Errantiviridae (blue), Caulimoviridae (red) and Athila/Tat (orange)

On the basis of these observations, we examined the intron size distribution among a larger set of orthologous genes involved in developmental processes. While introns of non-developmental genes in axolotl show a median size expansion of 13- to 25-fold compared to human, mouse and frog, the expansion of introns of developmental genes is significantly lower (6- to 11-fold, $P < 10^{-11}$) (Fig. 3a, Supplementary Information section 4.3). In contrast to human, mouse and frog, introns of developmental genes in axolotl are shorter than introns of non-developmental genes. Furthermore, axolotl multi-exon genes that contain only short introns exhibit gene ontology enrichments related to developmental patterning that are not enriched in multi-exon genes with larger introns (Supplementary Table 16). These results suggest that intron size in developmental genes is under constraint in the axolotl, possibly because smaller gene sizes facilitate rapid transcription and thus upregulation of these genes in specific developmental contexts.

A reduced Pax-family complement

Next, we interrogated the genome for families of canonical developmental signalling molecules (Supplementary Information section 5). All three hedgehog paralogs as well as a full set of vertebrate Wnt genes were present (Extended Data Fig. 2a, b). However, we noted that certain members of the paired box family of transcription factors, which have diverse roles in tissue formation, were not found in the assembly. Consistent with the absence of *Pax4* in amphibians and other vertebrate lineages¹⁷, the axolotl genome does not contain *Pax4* but does contain *Pax10*. Notably, despite the presence of the *Pax3* and *Pax7* paralogs in all other known vertebrate lineages, we were able to identify *Pax7* but not *Pax3* in the axolotl genome assembly (Extended Data Fig. 2c). No *Pax3* sequence was found in either the raw PacBio sequencing reads or the transcriptome. To confirm the loss of *Pax3*, we further examined the genomic region that would be syntenic for *Pax3* for the presence of neighbouring genes and highly-conserved non-exonic elements (CNEs). The orthologues of genes surrounding mouse *Pax3* (*Sgpp2* and *Epha4*) were present in the *A. mexicanum* genome assembly; however, neither the *Pax3* gene nor any of the *Pax3*-associated CNEs were found (Fig. 3d). By contrast, several CNEs that overlap the *Pax7* gene were identified in the assembly. Together, this evidence strongly suggests that *Pax3* and several of its *cis*-regulatory

families are not found. **c**, Box plots show the length distribution of LTR families (ERVL, endogenous retrovirus-like). Boxes indicate the first quartile, the median and the third quartile with whiskers extending up to 1.5 times the interquartile distance. Outliers are defined as data points outside the whiskers and are shown as dots. Quantitative data and sample sizes are shown in Source Data. **d**, Relative age (Kimura distance) suggests prolonged transposition activity followed by a recent activity burst.

elements are absent in the axolotl genome, probably owing to a deletion.

Axolotl *Pax7* has similar functions to *Pax3*

To functionally assess the consequence of the absence of *Pax3* in the axolotl, we used TALEN- and CRISPR-mediated gene editing¹⁸ to mutate *Pax7*. In other vertebrates, *Pax3* and *Pax7* play key roles in muscle, neural tube and neural crest-derived tissue development¹⁹. Although these two genes share some common functions, deletion of *Pax3* or *Pax7* causes distinct phenotypes in mice^{20–22}. We investigated whether frameshift deletions introduced into the *AmPax7* gene would yield a comparable *Pax7* phenotype, or whether *AmPax7* may have taken on functions that are carried out by *Pax3* in other vertebrates. Two different TALEN-mutant alleles (7-nt and 20-nt deletions) of *AmPax7* were bred through two generations (Fig. 4a, Supplementary Information section 6). In the F2 generation, the developmental phenotype described below was observed in 83 out of 345 (24%) progeny from the *Pax7*^{Δ20nt/+} intercrossing and 57 of 232 (24.6%) progeny from the *Pax7*^{Δ7nt/+} intercrossing (Fig. 4b, Extended Data Fig. 3). The phenotype was evident in homozygous mutants, as analysed by PCR and loss of protein (Supplementary Information sections 6.1, 6.3). This information, combined with the CRISPR-mediated gene mutation results (Supplementary Information sections 6.2), shows that the homozygous *Pax7*^{Δ20nt/Δ20nt} and *Pax7*^{Δ7nt/Δ7nt} mutants represent recessive, complete or partial loss of *Pax7* function.

The *Pax7*^{Δ20nt/Δ20nt} and strong F0 *Pax7*-CRISPR mutants exhibited a curved body, were unable to maintain an erect posture and exhibited a delay in growth compared to controls. Immunohistochemical analysis of trunk or tail cross-sections of early stage, 20-day-old *Pax7*^{Δ20nt/Δ20nt} or 17-day-old F0 *Pax7*-CRISPR axolotls showed normal muscle mass. However, at later stages, consistent with the mouse *Pax7* deletion phenotype, tail and trunk muscles were greatly decreased (Fig. 4c, Extended Data Figs 4–6). Remarkably, the *Pax7* mutant axolotls also completely lacked limb muscle (Figs 4d, Extended Data Fig. 7). In mice, *Pax3*, but not *Pax7*, is required for limb muscle formation^{21–23} (Supplementary Table 18). These results demonstrate that *AmPax7* has comparable functions to *MmPax3* in the control of limb muscle genesis.

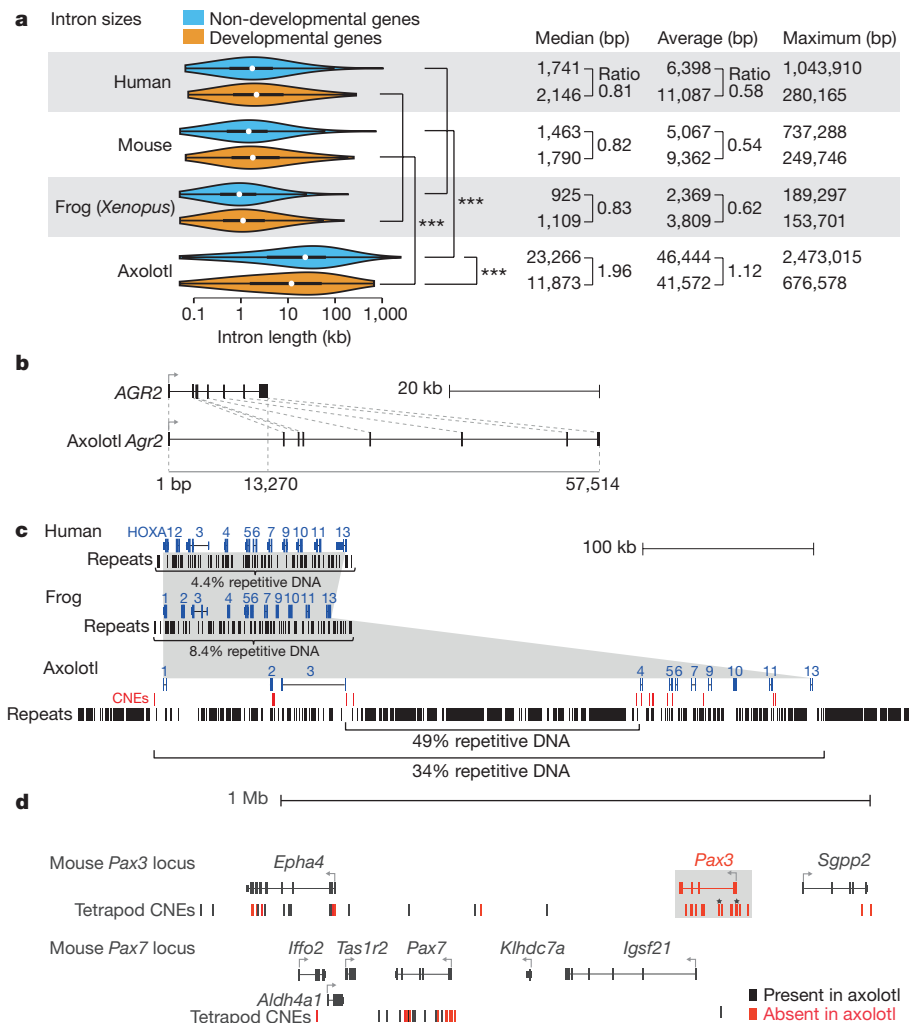


Figure 3 | Genome organization and loss of *Pax3*. **a**, Intron size of developmental genes appears to be under constraint in *A. mexicanum*. Violin plots represent the full distribution of intron sizes (thick bar, first to third quartile; white dot, median). *** $P < 10^{-11}$; two-sided Wilcoxon rank-sum tests. Quantitative data are shown in Source Data. **b**, Organization of the *Agr2* exon-intron structure shows a consistent expansion of axolotl intron sizes compared to those in the human

orthologue, resulting in a gene that is 4.3 times larger. **c**, Comparison of genes and repetitive elements in the HoxA cluster. CNEs that align to the axolotl HoxA cluster are shown in red. **d**, Axolotl lacks the *Pax3* locus. Analysis of tetrapod-conserved genes and CNEs associated with *Pax7* and *Pax3* genomic loci. Red, gene sequences and CNEs that are absent in axolotl; *CNEs that overlap well-characterized mouse *Pax3* enhancers^{38,39}.

In mice, *Pax7* deletion affects craniofacial neural crest derivatives, including the facial bones²⁰, whereas in zebrafish, *pax7* mutants show loss of xanthophores and reduction of melanophores, but no loss of iridophores²⁴. The Am*Pax7* mutants lacked a prefrontal bone, had a reduced number of melanophores and were deficient in xanthophores and iridophores except in the eyes (Fig. 4e–g, Extended Data Fig. 8). *Pax3* deletion in mice is associated with neural tube closure defects^{22,23} (Supplementary Table 18). Similarly, *Pax7* ^{$\Delta 20nt/\Delta 20nt$} -TALEN and *Pax7*-CRISPR axolotls displayed failed closure of the neural tube in the midbrain (Fig. 4h, Extended Data Fig. 9). In summary, mutation of Am*Pax7* yields a combination of the *Pax3*- and *Pax7*-mutant phenotypes that are observed in other vertebrates (Supplementary Information section 6). It will be interesting to understand how the regulation of *Pax7* has changed in axolotl to enable the loss of *Pax3*, which is essential in other vertebrates.

Species-restricted genes in regeneration

Previous searches for mRNA and microRNA (miRNA) transcripts associated with limb regeneration relied on mapping to *de novo* transcriptome assemblies. We sought to re-examine these datasets using our newly acquired genomic data. Recent functional work has highlighted the role of diverged gene or protein function during regeneration^{25–27}. Analysis of published tissue-enriched datasets²⁸,

combined with regeneration time courses^{29,30} and our own transcriptional profiling of 22 tissues, identified five transcripts that are upregulated in the limb blastema (the mass of proliferating cells involved in regenerating the limb) with orthology limited to non-amniote vertebrates (Supplementary Information section 7). One of these protein sequences shows a weak similarity to tectorin, a basement membrane component normally found in the inner ear, consistent with studies that implicate extracellular matrix components with having an important role in limb regeneration^{31,32}. Notably, another of these transcripts encodes a Ly6 family member in the urokinase type plasminogen activator surface receptor (uPAR) class. Previous studies had identified the salamander-specific Ly6 family member *Prod1* as a key factor involved in salamander limb development and regeneration^{25,33}. Our results suggest that Ly6 family members have a broader role in limb regeneration. Finally, we also investigated the role of non-coding RNAs by mapping a dataset of small RNA sequences expressed in the limb and limb blastema³⁴ to our genome assembly. This analysis classified 93 small RNAs as pre-miRNA sequences, of which 42 appear to be novel miRNAs (Supplementary Information section 7.2). Taken together, these data point to a potential role in limb regeneration for several coding and non-coding sequences that have been lost or diverged rapidly

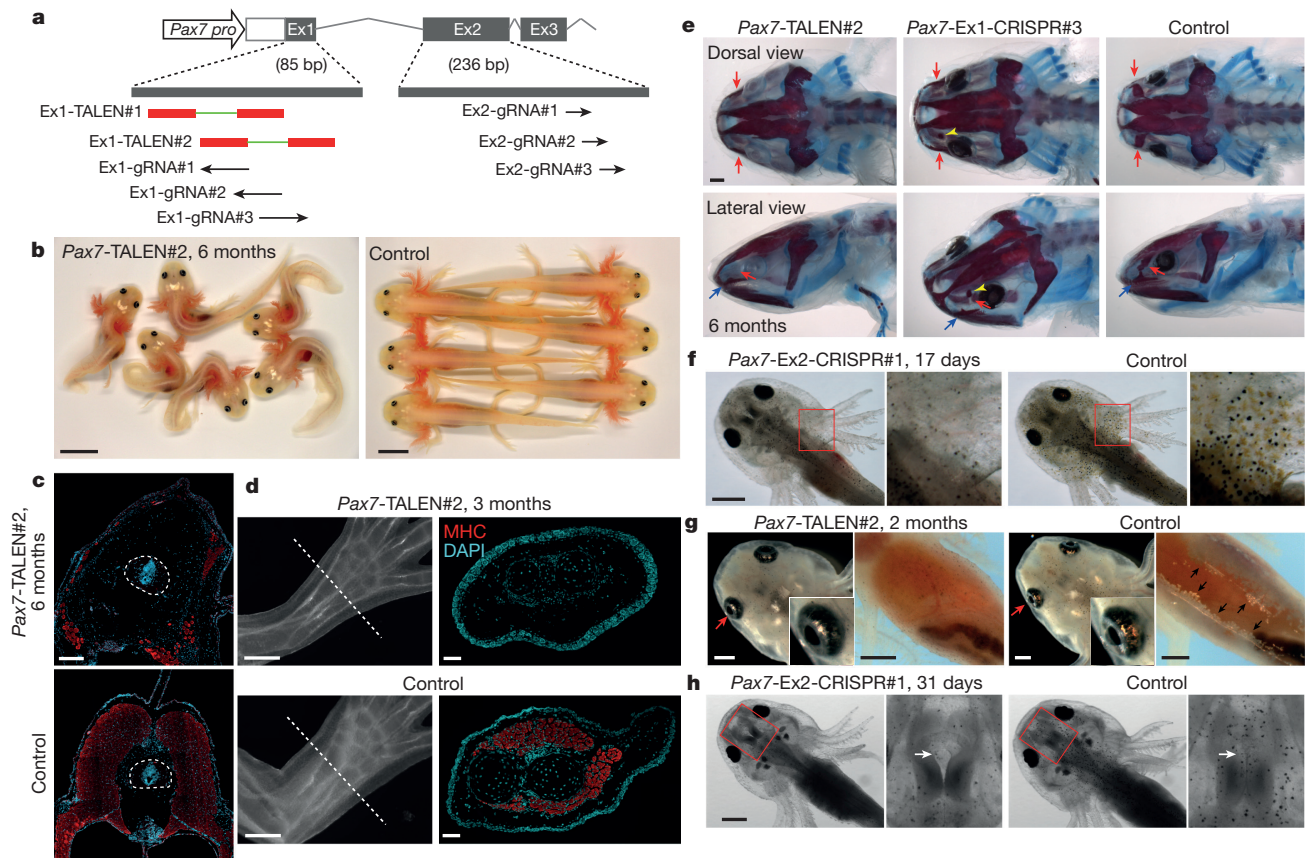


Figure 4 | Pax7 mutation in *A. mexicanum* yields a phenotype similar to that of Pax3^{-/-}Pax7^{-/-} mice. **a**, Deletion of AmPax7 coding sequences using TALEN and CRISPR. Deletions were made in exon 1 or exon 2. The first three AmPax7 exons (Ex) are shown. Red rectangles, TALEN targets; arrows, CRISPR-guide RNA (gRNA)-binding sites. **b**, Images of 6-month-old Pax7^{Δ20nt/Δ20nt} mutants compared to controls show loss of body elongation. Scale bars, 1 cm. **c**, Reduced body wall muscle in Pax7 mutants. Immunofluorescence images of myosin heavy chain (MHC, red) and DAPI (blue) in trunk cross-sections from a 6-month-old Pax7^{Δ20nt/Δ20nt} mutant (Pax7-TALEN#2) and a control animal. Scale bar, 500 μm. **d**, Limbs of Pax7 mutants lack muscle. Forelimb (left; scale bars, 500 μm) and immunofluorescence images of MHC (red) and DAPI (blue) in limb cross-sections (right; scale bars, 100 μm) of a 3-month-old Pax7^{Δ20nt/Δ20nt} mutant and a control animal. **e**, Loss of prefrontal bone in Pax7 mutants.

Dorsal and lateral views of Alcian blue and Alizarin red-stained Pax7^{Δ20nt/Δ20nt} and Pax7-Ex1-CRISPR#3 mutants and controls (right). Red arrows, prefrontal bone; blue arrows, maxillary bone. The yellow arrowhead points to a small remnant of the prefrontal bone. Scale bar, 1 mm. **f**, **g**, Reduced melanophores, xanthophores (**f**) and iridophores (**g**) in Pax7 mutants. Images of 17-day-old Pax7-Ex2-CRISPR#1 (**f**), 2-month-old Pax7^{Δ20nt/Δ20nt} (**g**) mutants and controls (Ctr). Right panels in **f** show a magnified view of the outlined area; red arrows in **g** point to the eyes that are magnified in the insets; black arrows indicate the belly iridophores in the control. Scale bars, 1 mm. **h**, Neural tube closure defect in Pax7 mutants. Images of a 31-day-old Pax7-Ex2-CRISPR#1 mutant and control (Ctr). Right, magnified view of the outlined area. Scale bar, 1 mm. Quantitative data and sample sizes are provided in the Life Sciences Reporting Summary and Source Data.

in amniotes. Future investigations of such sequences are likely to be a fruitful avenue for understanding the evolution of regeneration capabilities.

Discussion

We have generated a comprehensive whole-genome assembly for the salamander *A. mexicanum*, and analysis of this assembly has allowed us to draw conclusions about the structure of the expanded genome. Our data, together with data from plants and partial data from several other salamander species, show that LTR expansion is a major contributor to giant genome size across animals and plants^{6,12,35}. Our assembly is sufficiently complete to reliably detect the absence of Pax3, which is present in fish and other amphibians. This analysis was confirmed using gene editing, which showed that AmPax7 has assumed functions that are carried out by Pax3 in other animals.

Functional analysis of axolotl development, physiology and regeneration is facilitated by the availability of tissue- and time-dependent gene expression profiles^{28–30,36}. The axolotl genome provides a foundation for applying methods such as chromatin immunoprecipitation with sequencing (ChIP-seq) or assay for transposase-accessible chromatin using sequencing (ATAC-seq) to investigate the genomic basis of

gene regulation during regeneration. Together with methods such as CRISPR-mediated gene editing, viral expression methods, transplantation and transgenesis, the axolotl is a powerful system for studying questions such as the evolutionary basis of its remarkable regeneration ability. Our approach of long-read sequencing, optical mapping and genome assembly using MARVEL also demonstrates that it is now feasible to assemble very large repeat-rich genomes.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 April; accepted 13 December 2017.

Published online 24 January 2018.

- Spemann, H. & Mangold, H. Über Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren. *Arch. Mikrosk. Anat. En.* **100**, 599–638 (1924).
- Sperry, R. W. Effect of 180 degree rotation of the retinal field on visuomotor coordination. *J. Exp. Zool.* **92**, 263–279 (1943).
- Voss, S. R., Epperlein, H. H. & Tanaka, E. M. *Ambystoma mexicanum*, the axolotl: a versatile amphibian model for regeneration, development, and evolution studies. *Cold Spring Harb. Protoc.* 2009, <http://www.dx.doi.org/10.1101/pdb.em0128> (2009).

4. Currie, J. D. *et al.* Live imaging of axolotl digit regeneration reveals spatiotemporal choreography of diverse connective tissue progenitor pools. *Dev. Cell* **39**, 411–423 (2016).
5. Tanaka, E. M. The molecular and cellular choreography of appendage regeneration. *Cell* **165**, 1598–1608 (2016).
6. Keinath, M. C. *et al.* Initial characterization of the large genome of the salamander *Ambystoma mexicanum* using shotgun and laser capture chromosome sequencing. *Sci. Rep.* **5**, 16413 (2015).
7. Warren, R. L. *et al.* Improved white spruce (*Picea glauca*) genome assemblies and annotation of large gene families of conifer terpenoid and phenolic defense metabolism. *Plant J.* **83**, 189–212 (2015).
8. Zimin, A. V. *et al.* An improved assembly of the loblolly pine mega-genome using long-read single-molecule sequencing. *Gigascience* **6**, 1–4 (2017).
9. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
10. Sun, Y.-B. *et al.* Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes. *Proc. Natl Acad. Sci. USA* **112**, E1257–E1262 (2015).
11. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
12. Sun, C. & Mueller, R. L. Hellbender genome sequences shed light on genomic expansion at the base of crown salamanders. *Genome Biol. Evol.* **6**, 1818–1829 (2014).
13. Smith, J. J. *et al.* Genic regions of a large salamander genome contain long introns and novel genes. *BMC Genomics* **10**, 19 (2009).
14. Davis, A. P., Witte, D. P., Hsieh-Li, H. M., Potter, S. S. & Capocchi, M. R. Absence of radius and ulna in mice lacking *hoxa-11* and *hoxd-11*. *Nature* **375**, 791–795 (1995).
15. Roensch, K., Tazaki, A., Chara, O. & Tanaka, E. M. Progressive specification rather than intercalation of segments during limb regeneration. *Science* **342**, 1375–1379 (2013).
16. Voss, S. R. *et al.* Salamander Hox clusters contain repetitive DNA and expanded non-coding regions: a typical Hox structure for non-mammalian tetrapod vertebrates? *Hum. Genomics* **7**, 9 (2013).
17. Feiner, N., Meyer, A. & Kuraku, S. Evolution of the vertebrate *Pax4/6* class of genes with focus on its novel member, the *Pax10* gene. *Genome Biol. Evol.* **6**, 1635–1651 (2014).
18. Gaj, T., Gersbach, C. A. & Barbas, C. F. 3rd ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).
19. Mansouri, A., Hallonet, M. & Gruss, P. Pax genes and their roles in cell differentiation and development. *Curr. Opin. Cell Biol.* **8**, 851–857 (1996).
20. Mansouri, A., Stoykova, A., Torres, M. & Gruss, P. Dysgenesis of cephalic neural crest derivatives in *Pax7*^{-/-} mutant mice. *Development* **122**, 831–838 (1996).
21. Kuang, S., Chargé, S. B., Seale, P., Huh, M. & Rudnicki, M. A. Distinct roles for *Pax7* and *Pax3* in adult regenerative myogenesis. *J. Cell Biol.* **172**, 103–113 (2006).
22. Relaix, F., Rocancourt, D., Mansouri, A. & Buckingham, M. Divergent functions of murine *Pax3* and *Pax7* in limb muscle development. *Genes Dev.* **18**, 1088–1105 (2004).
23. Auerbach, R. Analysis of the developmental effects of a lethal mutation in the house mouse. *J. Exp. Zool.* **127**, 305–329 (1954).
24. Nord, H., Dennhag, N., Muck, J. & von Hofsten, J. *Pax7* is required for establishment of the xanthophore lineage in zebrafish embryos. *Mol. Biol. Cell* **27**, 1853–1862 (2016).
25. da Silva, S. M., Gates, P. B. & Brockes, J. P. The newt ortholog of CD59 is implicated in proximodistal identity during amphibian limb regeneration. *Dev. Cell* **3**, 547–555 (2002).
26. Garza-Garcia, A. A., Driscoll, P. C. & Brockes, J. P. Evidence for the local evolution of mechanisms underlying limb regeneration in salamanders. *Integr. Comp. Biol.* **50**, 528–535 (2010).
27. Sugiura, T., Wang, H., Barsacchi, R., Simon, A. & Tanaka, E. M. MARCKS-like protein is an initiating molecule in axolotl appendage regeneration. *Nature* **531**, 237–240 (2016).
28. Bryant, D. M. *et al.* A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. *Cell Reports* **18**, 762–776 (2017).
29. Stewart, R. *et al.* Comparative RNA-seq analysis in the unsequenced axolotl: the oncogene burst highlights early gene expression in the blastema. *PLoS Comput. Biol.* **9**, e1002936 (2013).
30. Knapp, D. *et al.* Comparative transcriptional profiling of the axolotl limb identifies a tripartite regeneration-specific gene program. *PLoS ONE* **8**, e61352 (2013).
31. Calve, S., Odelberg, S. J. & Simon, H.-G. A transitional extracellular matrix instructs cell behavior during muscle regeneration. *Dev. Biol.* **344**, 259–271 (2010).
32. Tassava, R. A., Nace, J. D. & Wei, Y. Extracellular matrix protein turnover during salamander limb regeneration. *Wound Repair Regen.* **4**, 75–81 (1996).
33. Kumar, A., Gates, P. B., Czarkwani, A. & Brockes, J. P. An orphan gene is necessary for preaxial digit formation during salamander limb development. *Nat. Commun.* **6**, 8684 (2015).
34. King, B. L. & Yin, V. P. A conserved microRNA regulatory circuit is differentially controlled during limb/appendage regeneration. *PLoS ONE* **11**, e0157106 (2016).
35. Sun, C. *et al.* LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol. Evol.* **4**, 168–183 (2012).
36. Jiang, P. *et al.* Analysis of embryonic development in the unsequenced axolotl: waves of transcriptomic upheaval and stability. *Dev. Biol.* **426**, 143–154 (2017).
37. Llorens, C. *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* **39**, D70–D74 (2011).
38. Milewski, R. C. *et al.* Identification of minimal enhancer elements sufficient for *Pax3* expression in neural crest and implication of *Tead2* as a regulator of *Pax3*. *Development* **131**, 829–837 (2004).
39. Degenhardt, K. R. *et al.* Distinct enhancers at the *Pax3* locus can function redundantly to regulate neural tube and neural crest expressions. *Dev. Biol.* **339**, 519–527 (2010).

Supplementary Information is available in the online version of the paper.

Acknowledgements E.W.M. was supported by the Max Planck Society President's fund and BMBF grants 01IS14014C and 031L0102. E.M.T., S.N. and J.-F.F. were supported by MPI-CBG, DFG FZ111 and IMP. E.M.T. and J.-F.F. were supported by DFG TA 274/3-3. J.-F.F. is supported by NSFC grant 31771611 and a Research Starting Grant S82111-E2 from SCNU. F.F. was supported by a CONACYT Masters Fellowship. A.C. was supported by The Swedish IRLG GRANT 2014-9040-114152-32, CONACYT FOINS-301 and Ciencia Básica IO017-CB-2015-01-00000000252126. G.Y. was supported by the Crick Institute award FC001162 (to J. P. Stoye). The Crick receives its core funding from CRUK, the MRC and the Wellcome Trust. B.H. was supported by BMBF grant 01IH11003C. We thank W. Bonacci, E. Gromberg, A. Tazaki and I. Stützer for sample preparation; S. Clausen, N. Gscheidel, Y. Duport, A. Sommer and the VBCF NGS facility for sequencing; A. Kavirayani and T. Engelmaier at VBCF for histology; E. Heude for confirmation of *Pax7* mutant immunofluorescence; T. Anantharaman for optimizing assembly and scaffolding parameters for Bionano data; and G. Papoutsoglou for advice.

Author Contributions S.N. assembled and analysed the transcriptome, and performed genome analysis. S.S. created, engineered and implemented the assembly algorithm and analysed the genome assembly. J.-F.F. performed DNA extraction and all biological experiments and analysis. M.P. and S.P. contributed to implementing the assembler, assembled the genome and contributed to the genome analysis. A.D. and S.W. performed Pacific Biosciences sequencing. G.Y. analysed the LTR elements. J.G.R. performed the conserved element and intron analysis. F.F. and A.C. performed the developmental orthologue and miRNA analysis. A.W.C.P., A.R.H. and H.C. performed Bionano optical mapping, generated the hybrid and scaffolded the assembly. D.K. performed tectorin analysis. B.H. supervised transcriptome assembly and annotation and acquired funding. E.M.T. and M.H. conceived analytical strategies, performed data analysis and acquired funding. S.N., S.S., J.-F.F., M.P., S.P., E.M.T. and M.H. wrote the manuscript. E.W.M. conceived and implemented the assembly strategy, acquired major funding and edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.H. (hiller@mpi-cbg.de), E.M.T. (elly.tanaka@imp.ac.at) and S.S. (siegfried.schloissnig@h-its.org).

Reviewer Information Nature thanks C. Crews and the other anonymous reviewer(s) for their contribution to the peer review of this work.



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Axolotl genomic DNA was prepared from freshly isolated liver and spleen of an individual three year old adult D/D male using DNAzol followed by phenol/chloroform extraction and ethanol precipitation.

A total of 50 size-selected SMRTbell libraries were prepared with a minimum fragment length cutoff between 10 kb and 20 kb. We sequenced medium and large insert libraries on the PacBio RSII instrument, making use of three different sequencing polymerases (P4, P5 and P6) and the corresponding sequencing chemistries (C2, C3 and C4). Movie times ranged from 180 min to 360 min with the majority of SMRT cells (1,414 of 1,933) at 240 min.

Sequences were assembled using the MARVEL assembler.

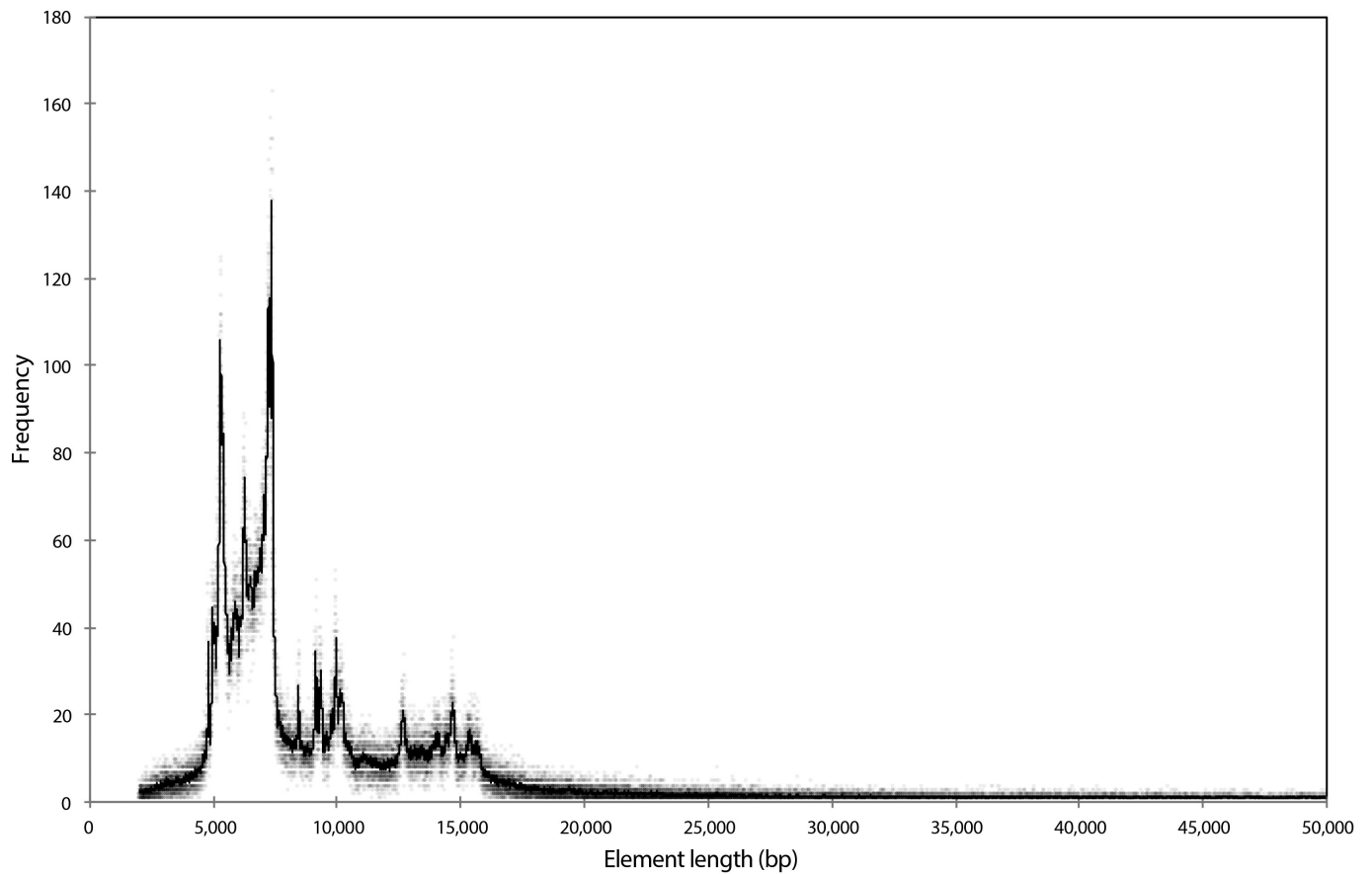
Optical mapping was performed using the Saphyr System (Bionano) based on NanoChannel array Technology. DNA was labelled with Nt.BspQI and Nb.BssSI enzymes in separate labelling reactions. Each enzyme reaction was run on the

Saphyr System. 2.813 Tb of data were collected on three Saphyr Chips for Nt.BspQI and 2.0 Tb of data were collected on two Saphyr Chips for Nb.BssSI samples; single molecule N50 lengths were 240 kb and 184 kb, respectively. Each dataset was *de novo* assembled using Bionano Solve 2.1 software.

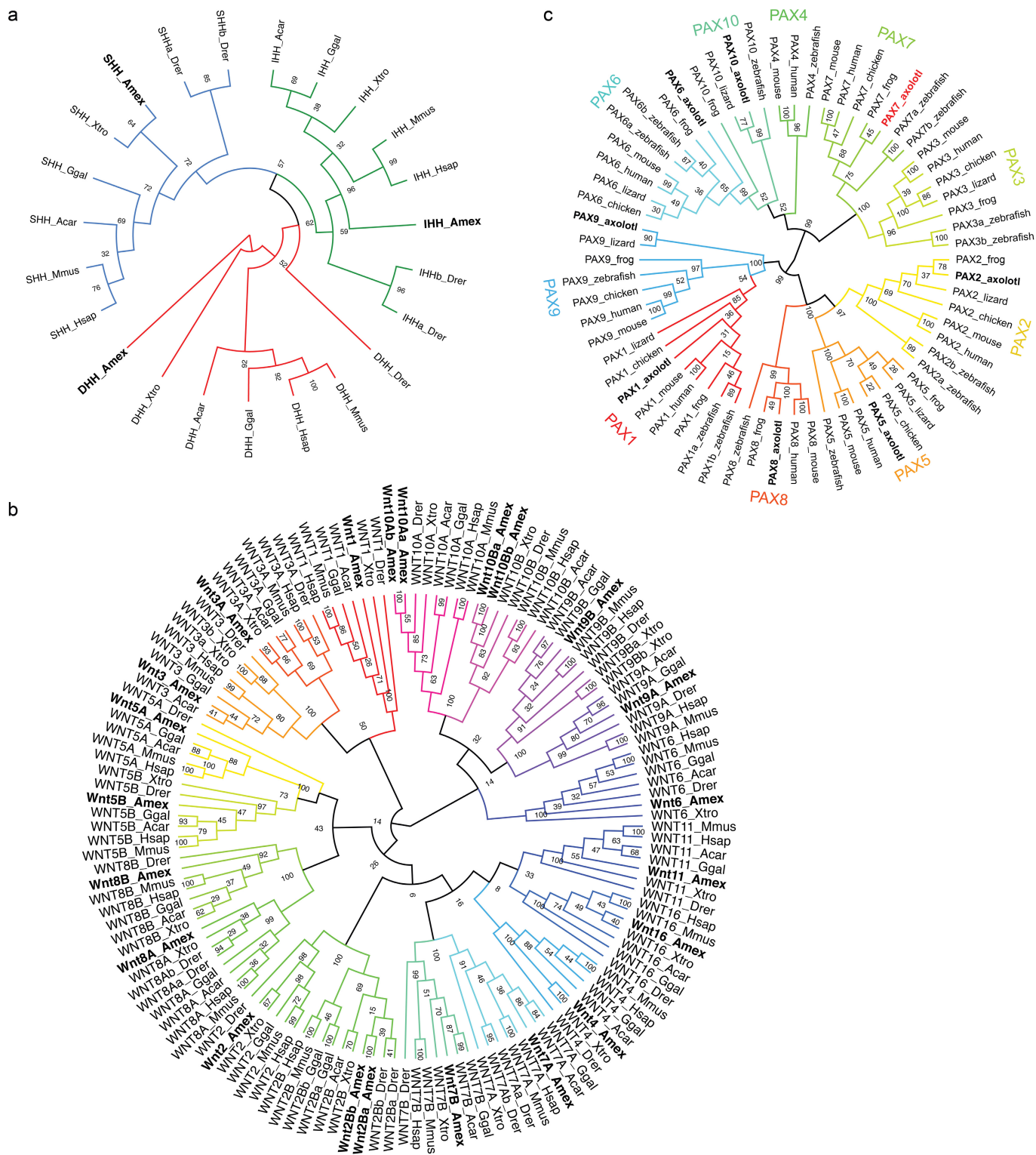
RNA was isolated from 22 tissue types using TRIzol or RNeasy reagents and sequenced using Illumina technology. The Trinity software package was used for transcriptome assembly.

Code availability. The MARVEL assembler with documentation is available at <https://github.com/schloi/MARVEL>.

Data availability. A browser of the axolotl genome is available at <https://genome.axolotl-omics.org>. The transcriptome assembly and the genome and transcriptome BLAST database can be accessed at <https://www.axolotl-omics.org> with no restrictions. The sequence data and both assemblies have been deposited in the NCBI BioProject database with accession numbers PRJNA378970 (genome data) and PRJNA378982 (transcriptome data). Both genome data and transcriptome data were deposited to the NCBI Nucleotide Database (nucleotide) with accession numbers PGSH000000000 and GFZP000000000, respectively.

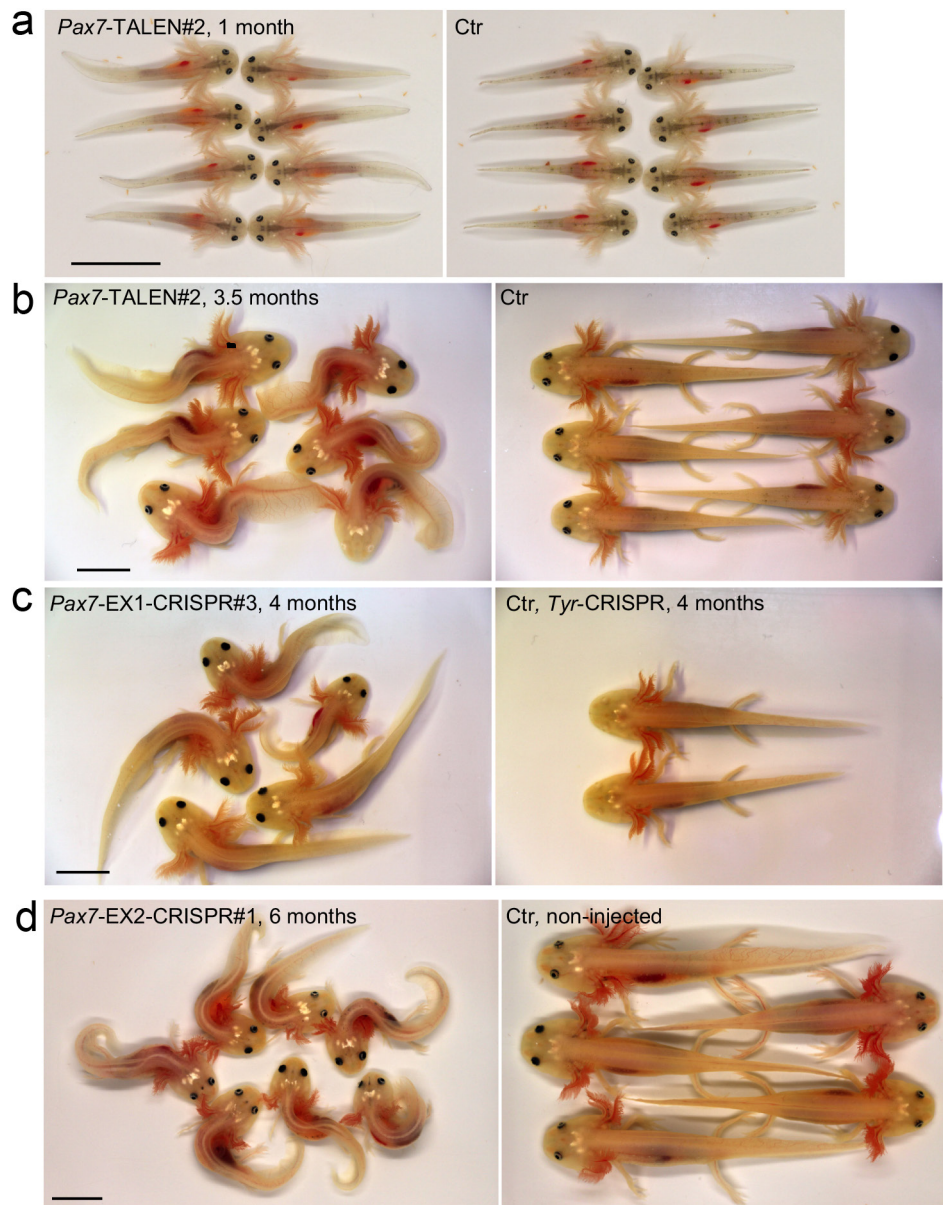


Extended Data Figure 1 | Analysis of LTR retroelement frequencies according to their lengths. The line shows a moving average (period 25) to highlight clusters of elements of similar lengths.



Extended Data Figure 2 | Phylogenetic trees. **a**, Phylogenetic tree of vertebrate hedgehog proteins show the presence of axolotl orthologues. **b**, Phylogenetic tree of vertebrate Wnt proteins show the presence of

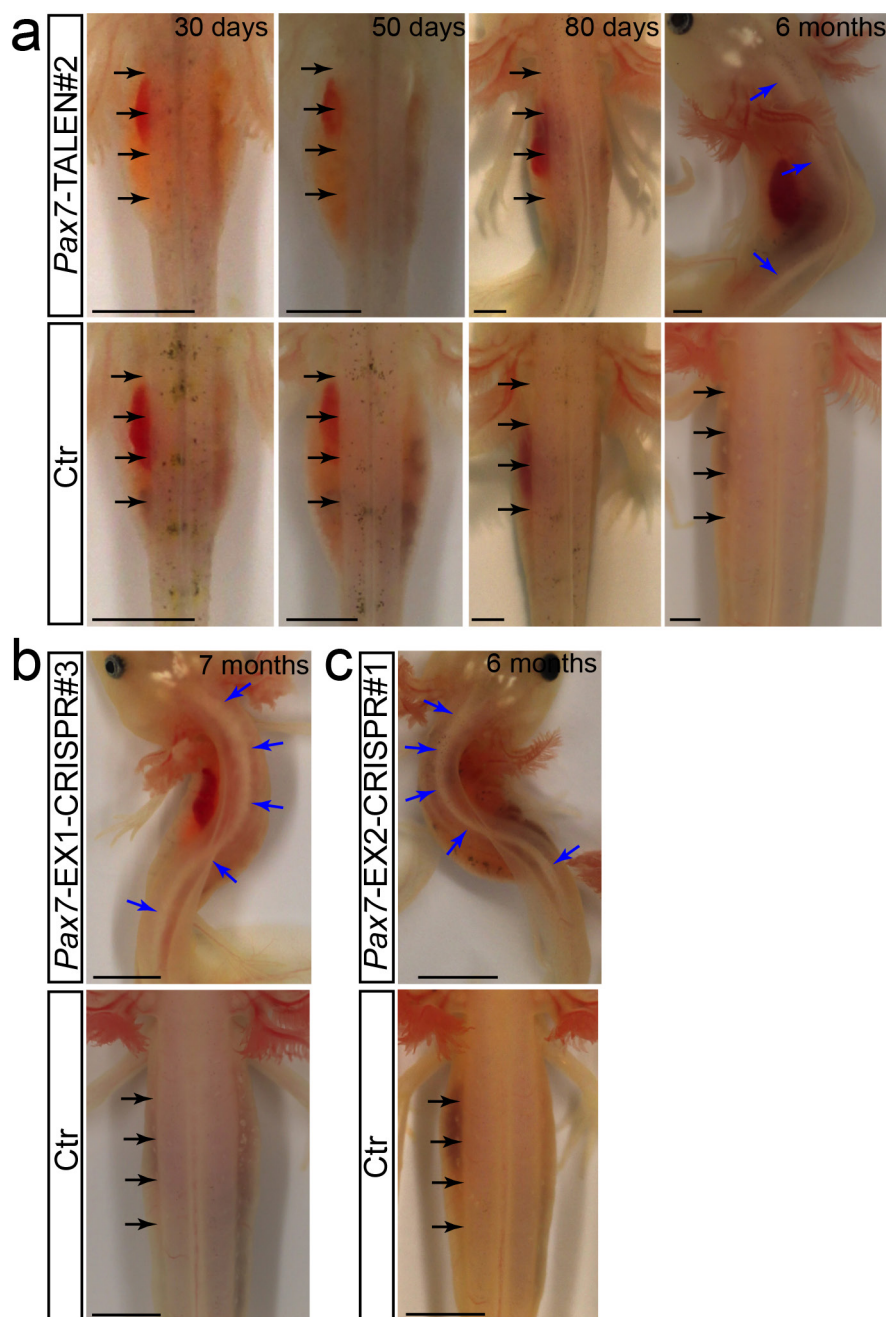
axolotl orthologues in all Wnt classes. **c**, Phylogenetic tree of vertebrate PAX proteins. *Pax4* and *Pax3* are absent in axolotl.



Extended Data Figure 3 | Developmental phenotype of *Pax7* mutants.

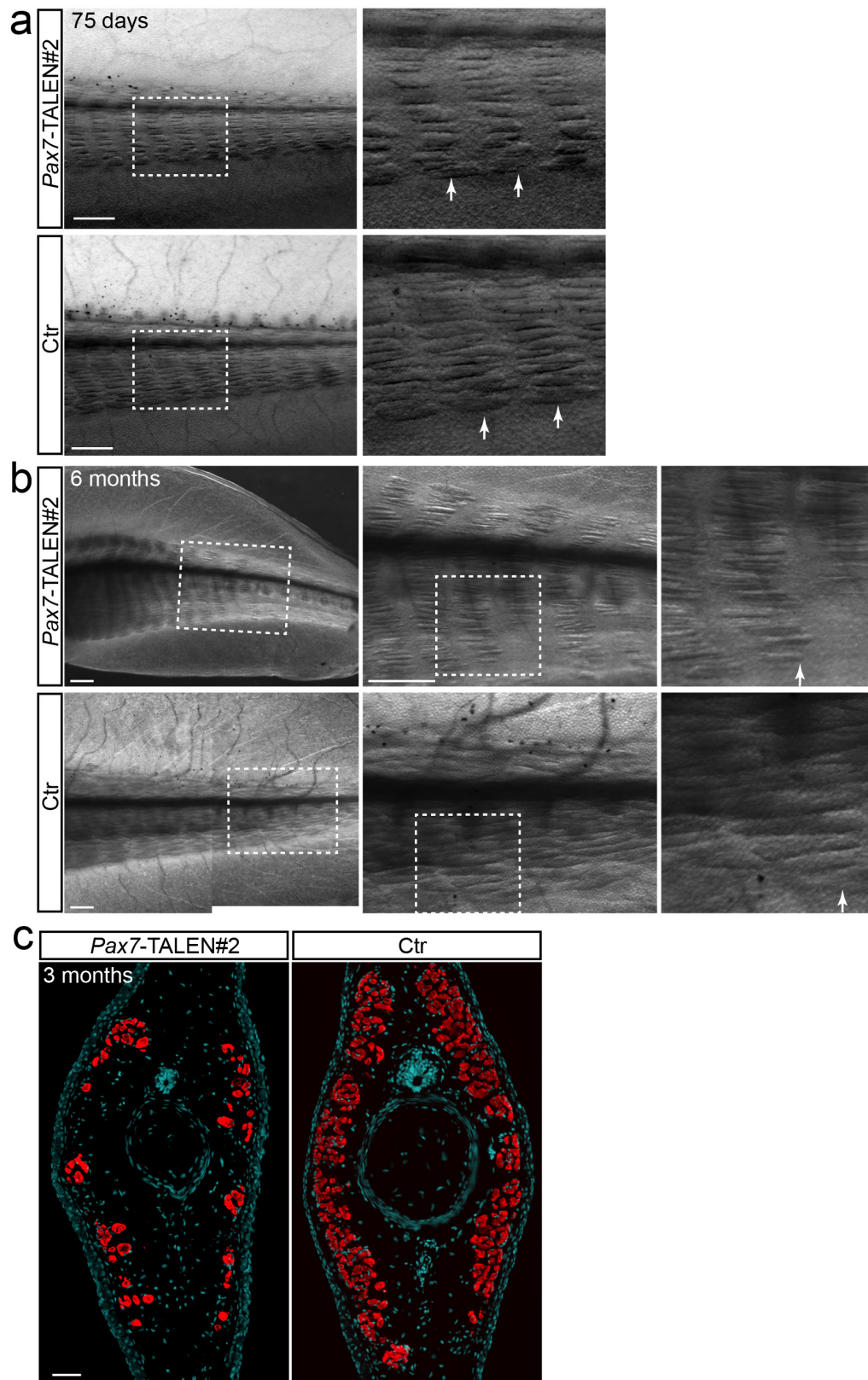
a, b, Images of live *Pax7* ^{$\Delta 20nt/\Delta 20nt$} mutants compared to controls show no obvious phenotype at early stages (**a**, 1 month), but an obvious phenotype at later stages (**b**, 3.5 months). **c, d**, Images of live F0 *Pax7*-Ex1-CRISPR#3

(**c**, 4 months) and *Pax7*-Ex2-CRISPR#1 (**d**, 6 months) mutants show the curved body phenotype. Scale bars, 1 cm. Numbers of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



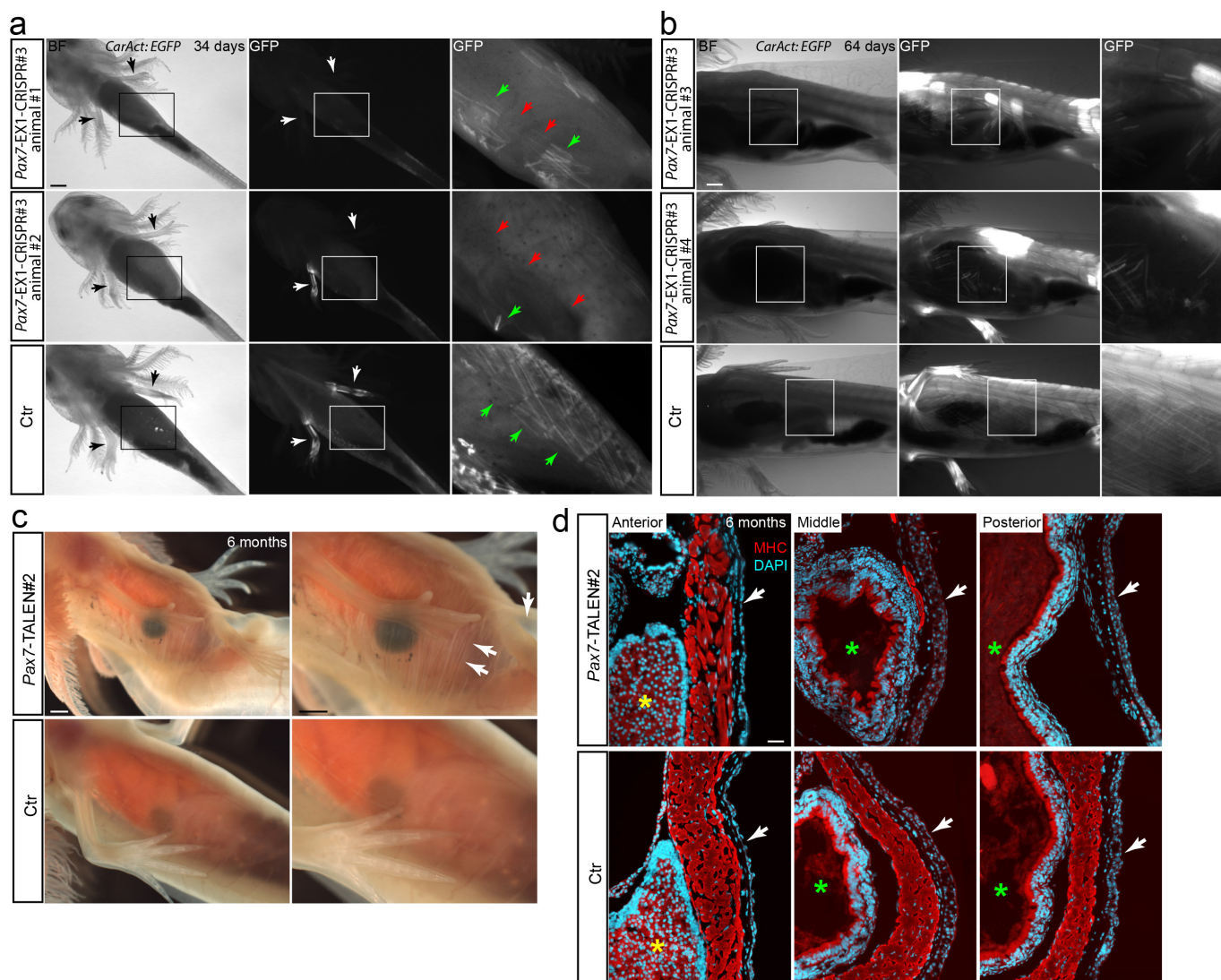
Extended Data Figure 4 | Progressive depletion of the trunk muscle in *Pax7* mutants. **a**, Images of live *Pax7* ^{$\Delta 20nt/\Delta 20nt$} animals at different ages compared to corresponding controls show the progressive loss of the trunk muscle in mutant animals. Black arrows indicate trunk muscles; blue arrows highlight the visibility of the spine after reduction and/or depletion of trunk muscle. Scale bars, 2mm. **b**, **c**, Images of live 7-month-old F0

Pax7-Ex1-CRISPR#3 (**b**) and 6-month-old F0 *Pax7*-Ex2-CRISPR#1 (**c**) mutants compared to controls, showing loss of trunk muscle. Black arrows indicate trunk muscles; blue arrows indicate the visibility of the spine after depletion of trunk muscle. Scale bars, 5 mm. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



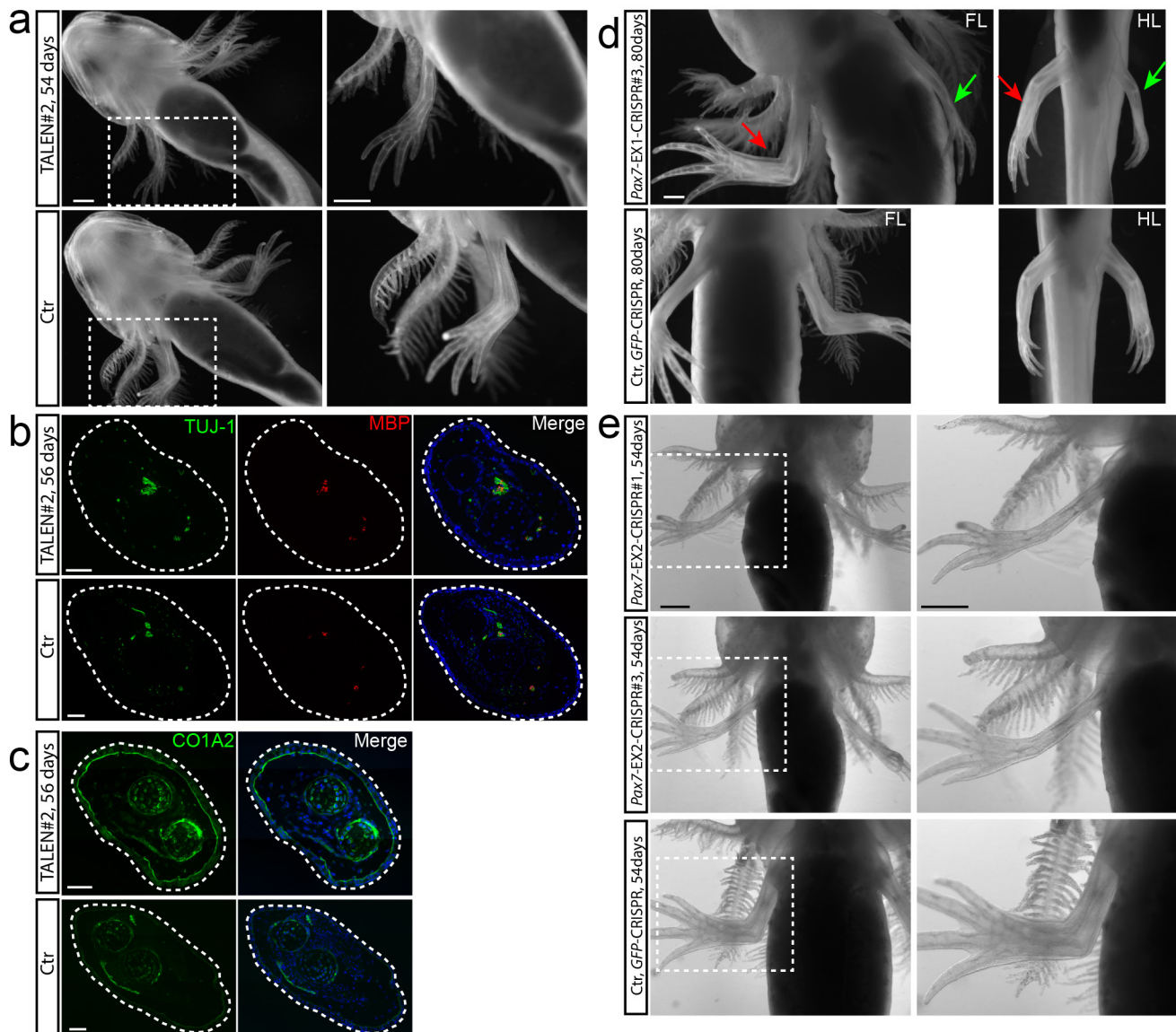
Extended Data Figure 5 | Progressive depletion of the tail muscle in *Pax7*^{Δ20nt/Δ20nt} mutants. **a, b,** Images of live 75-day (a) and 6-month-old (b) *Pax7*^{Δ20nt/Δ20nt} homozygous mutants compared to controls show the progressive depletion of tail muscle. White arrows indicate tail muscle fibres; right, magnified view of the outlined area. Note the decrease in myotome length in 75-day-old *Pax7*^{Δ20nt/Δ20nt} homozygous mutants (b).

Scale bars, 500 μm. **c,** Immunofluorescence images of MHC (red) and DAPI (blue) in tail cross-sections show reduction in tail muscle in 3-month-old *Pax7*^{Δ20nt/Δ20nt} mutants compared to controls. Scale bar, 100 μm. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



Extended Data Figure 6 | Depletion of the abdominal muscle in *Pax7* mutants. **a, b**, Bright-field image and GFP fluorescence of live 34-day (**a**) and 64-day-old (**b**) F0 *Pax7-Ex1-CRISPR#3* mutants obtained by injecting *Pax7*-gRNA#3–CAS9 protein complex into eggs of *CarAct:EGFP* transgenic axolotls, compared to un-injected *CarAct:EGFP* controls. Mutant animals show a reduction in the EGFP-labelled abdominal muscles. Right, magnified view of GFP fluorescence in the outlined area; white arrows indicate forelimbs that either contain or lack EGFP-labelled muscles; green arrows indicate the GFP-labelled abdominal muscle; red arrows indicate regions that lack GFP-labelled abdominal muscle.

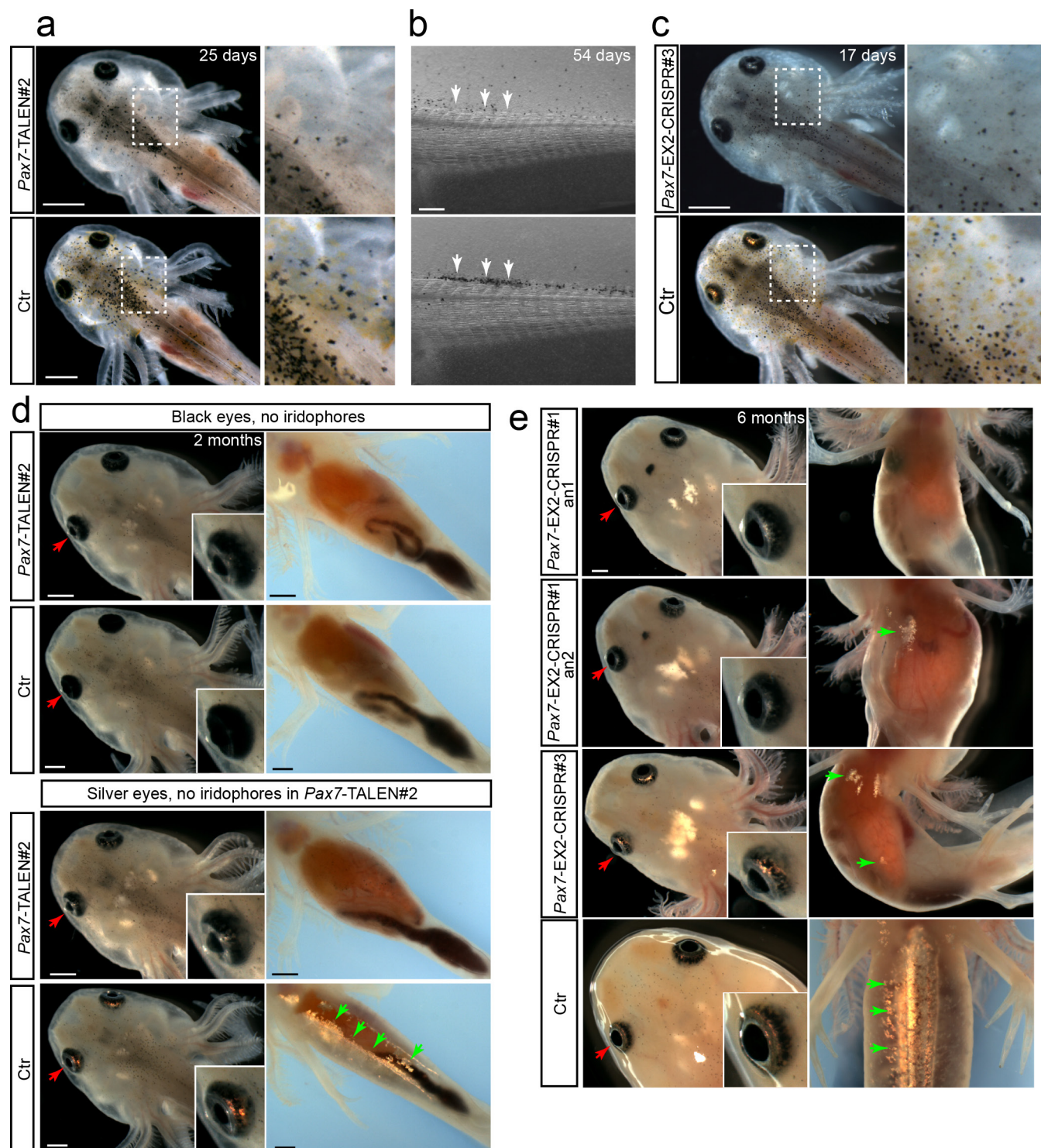
Scale bars, 1 mm. **c**, Images of live 6-month-old *Pax7*^{Δ20nt/Δ20nt} mutants compared to controls show the loss of abdominal muscle. Scale bars, 1 mm. **d**, Immunofluorescence images of MHC (red) and DAPI (blue) in cross-sections show the presence of ventral body-wall muscle in the chest position in 6-month-old *Pax7*^{Δ20nt/Δ20nt} homozygous mutants compared to controls, and the gradual depletion of the abdominal muscle along the anterior–posterior axis. Arrows, skin; yellow stars, liver; green stars, intestine. Scale bar, 100 μm. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



Extended Data Figure 7 | Loss of limb muscle in *Pax7* mutants.

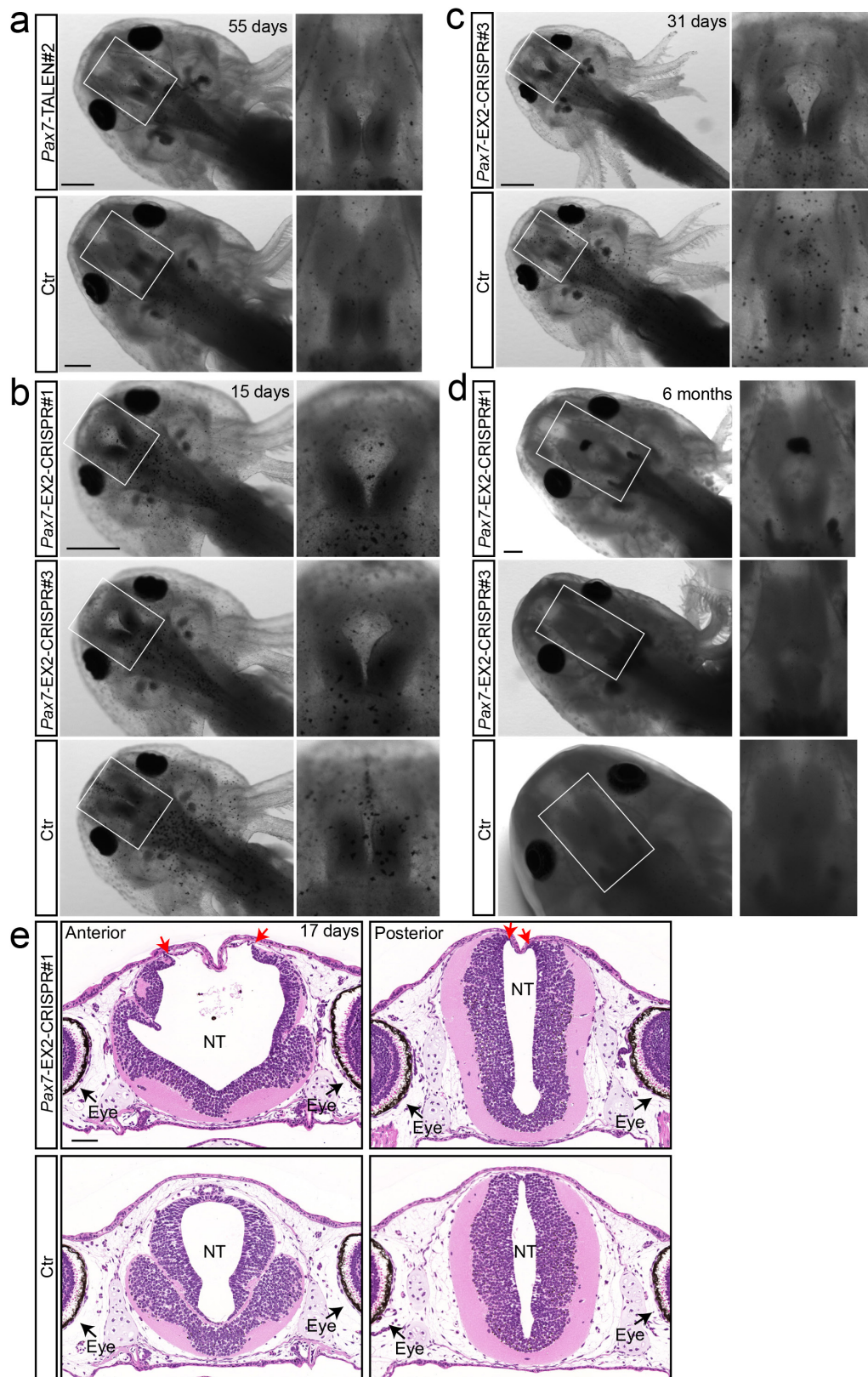
a, Images of live 54-day-old *Pax7*^{Δ20nt/Δ20nt} mutants compared to controls show loss of limb muscle. Right, magnified view of the outlined area. Scale bars, 1 mm. **b**, **c**, Non-muscle tissues are normal in *Pax7*^{Δ20nt/Δ20nt} mutant limbs. Immunofluorescence images for TUJ-1 (**b**, green) MBP (**b**, red), CO1A2 (**c**, green) and DAPI (blue) in forelimb cross-sections of 56-day-old *Pax7*^{Δ20nt/Δ20nt} mutants and controls. Scale bars, 100 μm. **d**, Images of live 80-day-old F0 *Pax7*-Ex1-CRISPR#3 heterozygotes compared to

controls show loss of forelimb (FL) and hindlimb (HL) muscle on one side of the body (green arrows) but not on the other side (red arrows). Scale bar, 1 mm. **e**, Images of live 54-day-old F0 *Pax7*-Ex2-CRISPR#1 and *Pax7*-Ex2-CRISPR#3 mutants compared to a control (bottom) showing loss of forelimb muscle in CRISPR animals. Right, magnified view of the outlined area. Scale bars, 1 mm. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



Extended Data Figure 8 | Reduced melanophores, loss of xanthophores and iridophores in *Pax7* mutants. **a**, Images of a live 25-day-old *Pax7*^{Δ20nt/Δ20nt} homozygous mutant compared to a control animal showing loss of xanthophores and reduction of melanophores in the head and neck region of mutant animals. Right, magnified view of the outlined area. Scale bar, 1 mm. **b**, Images of a live 54-day-old *Pax7*^{Δ20nt/Δ20nt} homozygous mutant compared to a control animal showing a reduction in melanophores along the body. Arrows, melanophores. Scale bar, 500 μm. **c**, Images of a live 17-day-old F0 *Pax7-Ex2-CRISPR#3* mutant compared to a control animal showing loss of xanthophores and reduction in melanophores in the head and neck region. Right, magnified view of the outlined area. Scale bar, 1 mm. **d**, Images of a live 2-month-old *Pax7*^{Δ20nt/Δ20nt} homozygous mutant compared to a control animal showing loss of iridophores on the belly. Red arrows point to the eye,

which is displayed at higher magnification showing eye pigmentation defects; green arrows indicate the presence of iridophores in the control animal (with silver eyes), but not in the mutant (with black eyes). Iridophores are absent in the *Pax7-TALEN#2* mutants, irrespective of the eye colour. Scale bars, 1 mm. **e**, Images of live 6-month-old *Pax7-Ex2-CRISPR#1* and *Pax7-Ex2-CRISPR#3* mutants compared to a control animal, showing the reduction or loss of belly iridophores (right) in axolotls with silver eyes (left). Red arrows point to the eye, which is displayed at a higher magnification on the right; green arrows indicate remaining iridophores in F0 mosaic *Pax7-CRISPR* mutants or iridophores in the control animal. Scale bars, 1 mm. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.



Extended Data Figure 9 | Neural tube closure defects in *Pax7* mutants.

a, Images of a live 55-day-old *Pax7* ^{Δ 20nt/ Δ 20nt} mutant compared to a control animal show an open brain phenotype. Right, magnified view of the outlined area. Scale bar, 1 mm. **b–d**, Images of live 15-day (**b**), 31-day (**c**) and 6-month-old (**d**) *Pax7-Ex2-CRISPR#1* and *Pax7-Ex2-CRISPR#3* mutants compared to controls, showing an open brain phenotype. Right,

magnified view of the outlined area. Scale bar, 1 mm. **e**, Haematoxylin and eosin-stained paraffin cross-sections show the open neural tube of a 17-day-old F0 *Pax7-Ex2-CRISPR#1* mutant compared to a control. Red arrows indicate the boundaries of the opened neural tube (NT). Scale bar, 200 μ m. Number of replicate matings and experiments are shown in the Life Sciences Reporting Summary and Source Data.

The genome of *Schmidtea mediterranea* and the evolution of core cellular mechanisms

Markus Alexander Grohme^{1*}, Siegfried Schloissnig^{2*}, Andrei Rozanski¹, Martin Pippel², George Robert Young³, Sylke Winkler¹, Holger Brandl¹, Ian Henry¹, Andreas Dahl⁴, Sean Powell², Michael Hiller^{1,5}, Eugene Myers¹ & Jochen Christian Rink¹

The planarian *Schmidtea mediterranea* is an important model for stem cell research and regeneration, but adequate genome resources for this species have been lacking. Here we report a highly contiguous genome assembly of *S. mediterranea*, using long-read sequencing and a *de novo* assembler (MARVEL) enhanced for low-complexity reads. The *S. mediterranea* genome is highly polymorphic and repetitive, and harbours a novel class of giant retroelements. Furthermore, the genome assembly lacks a number of highly conserved genes, including critical components of the mitotic spindle assembly checkpoint, but planarians maintain checkpoint function. Our genome assembly provides a key model system resource that will be useful for studying regeneration and the evolutionary plasticity of core cell biological mechanisms.

Rapid regeneration from tiny pieces of tissue makes planarians a prime model system for regeneration. Abundant adult pluripotent stem cells, termed neoblasts, power regeneration and the continuous turnover of all cell types^{1–3}, and transplantation of a single neoblast can rescue a lethally irradiated animal⁴. Planarians therefore also constitute a prime model system for stem cell pluripotency and its evolutionary underpinnings⁵. The taxonomic clade Platyhelminthes ('flatworms') also includes parasitic lineages that have substantial effects on human health, such as blood flukes (*Trematoda*) and tape worms (*Cestoda*)⁶. Here, the phylogenetic position of planarians as free-living flatworms⁷ provides a reference point towards an understanding of the evolution of parasitism⁸.

Despite the modest genome sizes of planarians (mostly in the range of 1–2 gigabase pairs (Gb)), genome resources relating to these animals are limited. Although the model species *S. mediterranea* was sequenced by Sanger sequencing, even 11.6× coverage of around 600-bp Sanger reads yielded only a highly fragmented assembly (N50 19 kb)⁹. Recent high-coverage, short-read approaches yielded similarly fragmented assemblies^{10,11}. The high A–T content (about 70%) represents one known assembly challenge. Furthermore, standard DNA isolation procedures perform poorly on planarians, which has so far precluded the application of long-read sequencing approaches or BAC-clone scaffolding.

We here report a highly contiguous PacBio SMRT long-read sequencing¹² assembly of the *S. mediterranea* genome. Giant gypsy/Ty3 retroelements, abundant AT-rich microsatellites and inbreeding-resistant heterozygosity collectively provide an explanation for why previous short-read approaches were unsuccessful. We find a loss of gene synteny in the genome of *S. mediterranea* and other flatworms. In analysis of highly conserved genes, we find a loss of MAD1 and MAD2, suggesting a MAD1–MAD2-independent spindle assembly check point (SAC)^{13,14}. Our *S. mediterranea* genome assembly provides a resource for probing the evolutionary plasticity of core cell biological mechanisms, as well as the genomic underpinnings of regeneration and the many other phenomena that planarians expose to experimental scrutiny.

De novo long read assembly of the planarian genome

In preparation for genome sequencing, we inbred the sexual strain of *S. mediterranea* (Fig. 1a) for more than 17 successive sib-mating generations in the hope of decreasing heterozygosity. We also developed a new DNA isolation protocol that meets the purity and high molecular weight requirements of PacBio long-read sequencing¹² (Extended Data Fig. 1a–d, Supplementary Information S1, S2). We used MARVEL, a new long-read genome assembler developed for low complexity read data¹⁵ (Supplementary Information S3). An initial *de novo* MARVEL assembly of reads of more than 4 kb with approximately 60× genome coverage showed an improvement over the PacBio assembly tool (Canu¹⁶) and substantial improvements over existing *S. mediterranea* assemblies based on short read sequencing (Extended Data Table 1). We further made use of the Chicago/HiRise *in vitro* proximity ligation method¹⁷ for scaffolding (Extended Data Fig. 1e, Supplementary Information S4). The polished haplotype-filtered (see below) and error-corrected (Supplementary Information S5) *S. mediterranea* assembly consists of 481 scaffolds with an N50 length of 3.85 Mb (Extended Data Table 1).

To assess the quality of this genome assembly, we back-mapped a transcriptome of the sequenced strain (Supplementary Information S6) and found that more than 99% of transcripts were mapped, thus confirming that the assembly was both near-complete and accurate (Supplementary Information S7, Extended Data Fig. 1f, g). To assess the contiguity of the global assembly, we analysed structural conflicts between the MARVEL assembly and Chicago/HiRise scaffolding. Out of 51 such events across the 782.1 Mb of assembled genome sequence, only two represented unambiguous MARVEL assembly mistakes (Fig. 1b, Supplementary Information S4.3). Furthermore, high-stringency back-mapping of high-confidence cDNA sequences (Supplementary Information S7.3) confirmed assembly contiguity below the approximately 1-kb resolution limit of the Chicago/HiRise method, with small-scale sequence duplications near assembly gaps as only minor inconsistencies (Extended Data Fig. 2).

Our *S. mediterranea* genome assembly represents a major improvement over existing *S. mediterranea* assemblies¹⁰ (Fig. 1c) and, to our

¹Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstraße 108, 01307 Dresden, Germany. ²Heidelberg Institute for Theoretical Studies, Schloss-Wolfsbrunnengasse 35, 69118 Heidelberg, Germany. ³The Francis Crick Institute, 1 Midland Road, London NW1 1AT, UK. ⁴Deep Sequencing Group, BIOTEC/Center for Regenerative Therapies Dresden, Cluster of Excellence at TU Dresden, Fetscherstraße 105, 01307 Dresden, Germany. ⁵Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Str. 38 01187 Dresden, Germany.

*These authors contributed equally to this work.

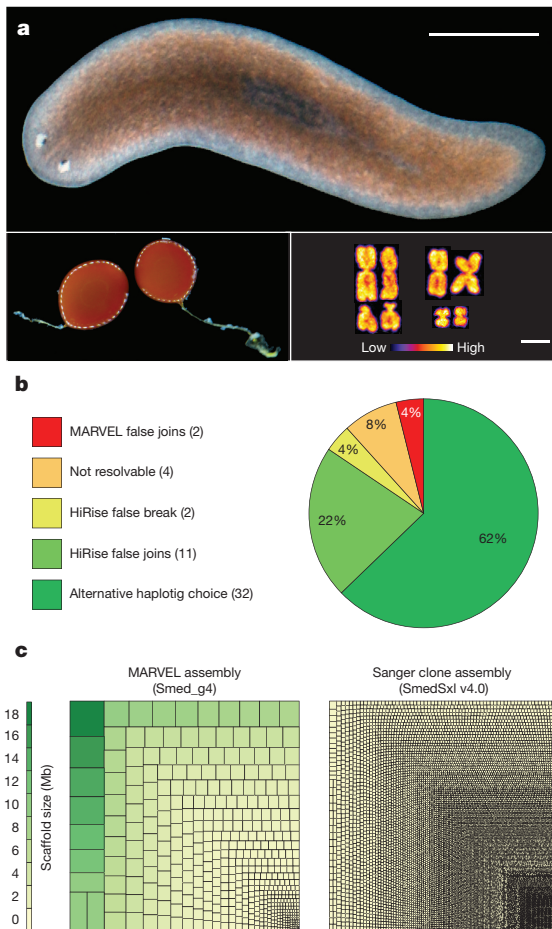


Figure 1 | Long-range contiguous genome assembly of *S. mediterranea*. **a**, Top, individual of the sequenced sexual strain. Bottom left, egg cocoons. Bottom right, karyotype (2N=8). Scale bars, 2 mm (top) and 2.5 μ m (bottom right). **b**, Chicago quality control of the assembly. **c**, Treemap comparison between the MARVEL *S. mediterranea* assembly and the most contiguous existing Sanger *S. mediterranea* assembly¹⁰. Squares encode the relative contributions of individual scaffolds or contigs to assembly size.

knowledge, is the first long-range contiguous assembly of the genome of a non-parasitic flatworm species. A UCSC genome browser instance with supplementary quality control, annotation and experimental data tracks (Supplementary Information S8) is available at PlanMine¹⁸ (<http://planmine.mpi-cbg.de>). All analyses in this manuscript refer to the assembly release version dd_Smed_g4. The current source code of the MARVEL assembler is available at <https://github.com/schloil/MARVEL>. The execution scripts used for *S. mediterranea* can be found in the smed subfolder of the examples folder.

Assembly challenges in the *S. mediterranea* genome

To understand why the *S. mediterranea* genome was recalcitrant to earlier short-read assembly, we first analysed its repeat content (Supplementary Information S9). The genome has a repetitive fraction of 61.7% (Fig. 2a), substantially exceeding the 38% or 46% repeat content of the mouse or human genomes, respectively¹⁹. We detected more than 7,000 insertions of 11 distinct families of long terminal repeat (LTR) retroelements (Fig. 2b, Extended Data Fig. 3a, Supplementary Information S10). These do not cluster with known *Metaviridae* (Fig. 2b), suggesting that they represent either extremely divergent or so far undescribed retroelement families. Three LTR families were more than 30 kb long—an exceptional size that is more than three times longer than the 5–10 kb typically observed in vertebrates (Fig. 2c, Extended Data Fig. 3b). The only known similar-sized LTRs are the plant-specific Ogre elements²⁰, which is why we refer to the giant

S. mediterranea repeat families as Burro (big, unknown repeat rivaling ogre; Supplementary Information S10.3). Burro elements are pervasively transcribed (Extended Data Fig. 3c, d, Supplementary Information S10.4), yet their high degree of intra-family sequence divergence suggests a relatively ancient invasion (Supplementary Table 1, Supplementary Information S10.5, Extended Data Fig. 3e). Burro-1, the most abundant giant retroelement with 130 fully assembled copies, is highly overrepresented at contig ends, and 50% of all current scaffolds terminate in a Burro-1 element (Fig. 2d, Supplementary Information S10.6). Therefore, these abundant, over 30-kb repeat elements still limit the size of the current assembly. In addition, abundant AT-rich microsatellite regions disrupt the alignment of spanning reads and thus also reduce contig continuity (Extended Data Fig. 4, Supplementary Information S11). Finally, the *S. mediterranea* assembly graphs showed substantial structural heterogeneity (Supplementary Information S12) in the form of bubbles (transient divergences in sequencing read alignments) and spurs (divergences without re-connection), which were largely absent from a comparable genome assembly (*Drosophila melanogaster* using PacBio sequencing and MARVEL assembly; Fig. 2e, Supplementary Information S12.1) or assemblies of 17 other species (Supplementary Table 2). Heterozygous mobile element insertions and microsatellite tracts were prominent causes of assembly divergences (Fig. 2f, Extended Data Fig. 4d, Supplementary Information S12.3). The persistence of substantial genomic heterozygosity in spite of more than 17 successive sib-mating generations confirms that meiotic recombination is inefficient in *S. mediterranea*²¹.

Overall, the combination of giant repeat elements, low-complexity regions and inbreeding-resistant heterozygosity provides an explanation for why previous short-read sequencing assemblies of *S. mediterranea* have proven so challenging. The long-range contiguity that we achieved in the *S. mediterranea* genome assembly, and similarly substantial improvements in the PacBio genome assembly of the flatworm *Macrostomum lignano*²² (Supplementary Table 2), further emphasize the improvements that a combination of long-read sequencing with the MARVEL assembler offers in the assembly of challenging genomes.

Comparative analysis of the planarian gene complement

We next annotated the *S. mediterranea* gene complement, relying on our planarian transcriptome resources¹⁸ (Supplementary Information S13). Our analysis showed a high divergence of *S. mediterranea* gene sequences (Supplementary Information S14), *en par* with *Caenorhabditis elegans* (Fig. 3a). By contrast, the low degree of sequence substitutions between the sexual and asexual *S. mediterranea* strains (Fig. 3a) and nearly identical mapping statistics of the two transcriptomes to the genome (Supplementary Information S7.1, Extended Data Fig. 1f) establish the utility of our assembly for both strains.

To evaluate the *S. mediterranea* genome structure, we performed whole-genome alignments (Supplementary Information S15) with the available parasitic flatworm genomes⁶ and a draft genome of the platyhelminth *M. lignano*²² (Fig. 3b). The highest alignment similarity was found between *S. mediterranea* and the parasitic flatworm *Schistosoma mansoni*, which is consistent with the platyhelminth phylogeny⁷. However, alignments were mostly limited to individual exons of specific genes, irrespective of the quality of the various assemblies (Extended Data Fig. 5a, b). In general, flatworm genome comparisons resulted in alignment chains that were much shorter and lower scoring than those obtained from comparisons across the tetrapod (human–frog) or vertebrate (human–zebrafish) clades (Fig. 3b). Together with more than 1,000 likely planarian-specific protein coding genes (Supplementary Information S16, Supplementary Table 5, Extended Data Fig. 6a–g), our data show a high degree of genome divergence between *S. mediterranea* and other flatworms.

We therefore next investigated gene loss in planarians. Our analysis deliberately focused on highly conserved genes, such that the absence of sequence similarity alone provides a strong indication of

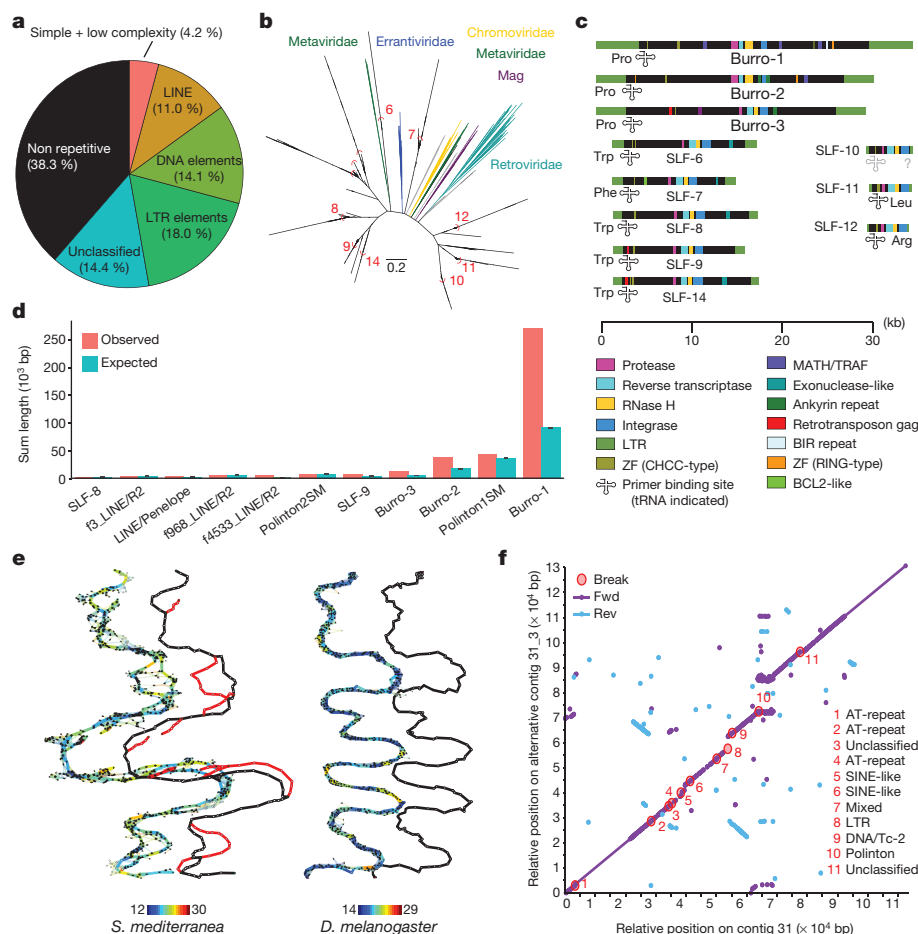


Figure 2 | *S. mediterranea* assembly challenges. **a**, Repeat content of the assembly. **b**, LTR family phylogeny. Known LTR families are shown in colour, *S. mediterranea* LTR families in black. Red arcs delimit clusters for consensus calculation. Scale bar: 0.2 substitutions per site. **c**, Domain annotation of the 11 *S. mediterranea* LTR families (SLFs). **d**, Enrichment analysis of indicated repeat elements within the terminal 1,000 bp of all scaffolds ($n = 962$). Expected values represent mean repeat frequency with 95% bootstrap confidence interval ($n = 1,000$). **e**, Graphical representation of representative *S. mediterranea* (left, ~ 1.6 Mb) and *D. melanogaster* (right, ~ 1.7 Mb) MARVEL PacBio assembly graph segments. Thick lines,

consensus sequence; thin lines, individual read alignments; colour-coding, alignment quality (blue, low; red, high; see spectra at bottom); black marks, repeats. The contig tour of the final haploid genome assembly is shown offset to the right and alternative regions are shown in red. **f**, Dot plot comparison between a representative alternative region and the corresponding main contig. Fwd, forward match; Rev, reverse match; Break, insertions or deletions over 99 bp. Break annotations (1–11, right) list repeat categories that cover more than 60% of the insertion/deletion sequence; ‘mixed’ indicates contributions of multiple repeat classes.

loss (Supplementary Information S17). We identified 452 highly conserved genes that were lost in both *S. mediterranea* and other planarians (Fig. 3c), which compares to 284 and 757 such losses in *D. melanogaster* and *C. elegans*, respectively (Extended Data Fig. 5c). Gene loss in planarians is therefore broadly in the same range as in established invertebrate model organisms. However, the lost genes included 124 homologues of genes that are essential in humans or mice (Supplementary Table 6) and are generally key components of multiple cell biological core mechanisms (Fig. 3c). Specifically, planarians lack multiple highly conserved components of DNA double-stranded break (DSB) repair, including *RAD52*, *XRCC4*, *NHEJ1* (also known as *XLF*), *SMC5*, *SMC6* and the entire condensin II complex²³. A possibly consequent reliance on mutagenic DSB repair pathways (for example, microhomology-mediated end joining)²⁴ could account for both the abundance of microsatellite repeats and the structural divergence of the *S. mediterranea* genome (Fig. 3b), but raises questions regarding the extraordinary resistance of planarians to DSB-inducing γ -irradiation⁴.

Planarians are also lacking recognizable homologues of key metabolic genes. Loss of the fatty acid synthase (*FASN*) gene is striking in the face of its essential role in *de novo* fatty acid synthesis in eukaryotes, and may indicate that planarians are particularly dependent on dietary lipids. The loss of the haem breakdown enzyme genes *HMOX1* and *BLVRB*

despite maintained haem biosynthesis capacity²⁵ is similarly unusual for a free-living eukaryote (both are lost in *C. elegans*²⁶). Remarkably, the above and multiple other genes were missing not only in planarians, but also in the parasite genomes⁶ and the transcriptome of the macrostomid *M. lignano*²⁷ (Fig. 3c). Given their broad conservation in the lophotrochozoan sister clade, the broad absence of these genes in flatworms is likely to represent an ancestral loss. This complicates, for example, the interpretation of *FASN* loss in the parasitic lineages as a specific adaptation to parasitism⁶. Conversely, the absence of key metabolic genes as phylogenetic signal underscores the utility of free-living flatworms as model systems for the parasitic lineages and the development of anti-helminth reagents⁸.

A MAD1–MAD2-independent spindle check-point?

The apparent absence of *MAD1* and *MAD2* in planarians (Fig. 3c) raises the question of whether planarians have a functional SAC, and how essential cellular functions can be maintained in the absence of supposed core components. Both *MAD1* and *MAD2* (also known as *MAD1L1* and *MAD2L1*) are near-universally conserved owing to their essential roles in the SAC, which guards against aneuploidy²⁸ by inhibiting cell cycle progression as long as even a single chromosome remains unattached to the mitotic spindle¹⁴.

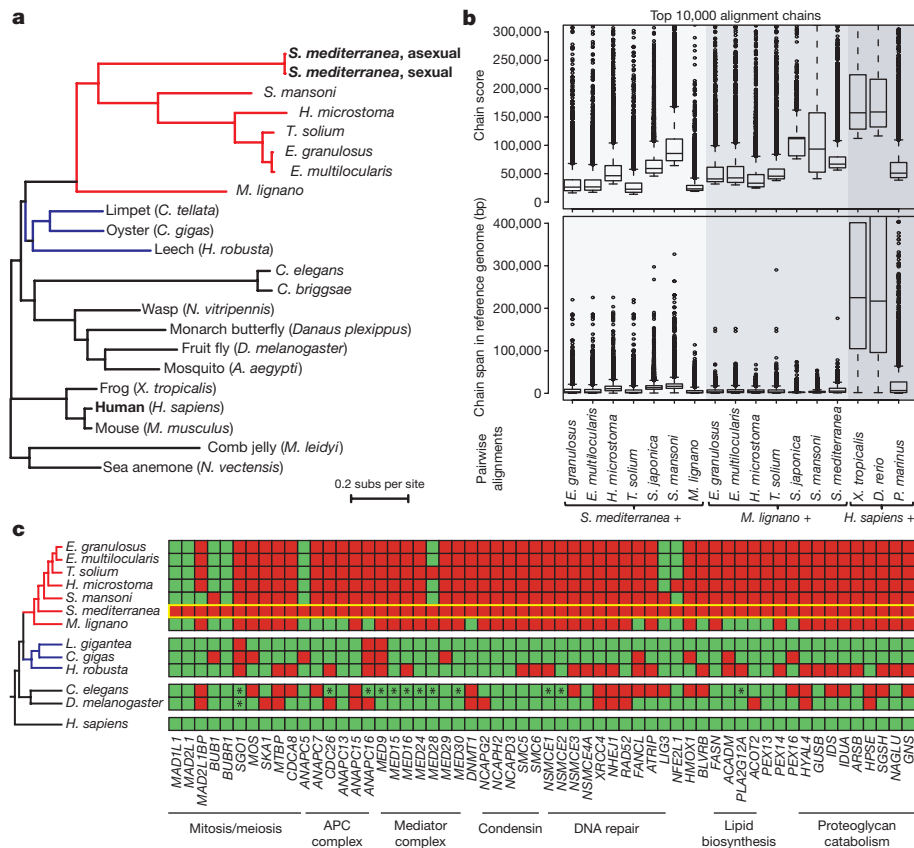


Figure 3 | Genome divergence of *S. mediterranea* and other flatworms. **a**, Protein sequence divergence amongst 51 single copy genes (Supplementary Table 3). Branch length shows substitutions per site. Red, flatworms; blue, lophotrochozoan outgroups. **b**, Whole genome alignments of *S. mediterranea*, *M. lignano* and *H. sapiens* against the indicated reference genomes. The distributions of the alignment score (top) and alignment span (bottom) of the top 10,000 chains of co-linear

alignments are shown as box plots, with boxes indicating the first quartile, median and third quartile with whiskers extending up to 1.5 times the interquartile distance. Outliers are defined as more than 1.5 times the interquartile and are shown as dots. **c**, Presence (green) or absence (red) of highly conserved genes in the indicated species. The yellow box highlights *S. mediterranea*. Asterisks mark homologues secondarily identified by manual searches.

Although MAD1 and MAD2 homologues are easily identifiable in all other flatworms examined (Extended Data Figs 7, 8), not even flatworm queries could identify significant homologues in *S. mediterranea* or the transcriptomes of five other planarian species. Therefore, planarians are likely to have lost MAD1, MAD2 and multiple other SAC components (Fig. 4a). The known M-phase arrest of planarian cells upon pharmacological interference with spindle function²⁹ (Fig. 4b) is therefore remarkable, as it indicates the maintenance of a SAC-like response despite a lack of supposed SAC core components.

To explore the underlying mechanisms of the SAC-like response in *S. mediterranea*, we targeted remaining components of the SAC network (Fig. 4a) by RNA interference (RNAi) and quantified the fraction of M-phase arrested cells with or without the microtubule depolymerizing drug nocodazole (Fig. 4b, Supplementary Information S18). The marked increase in the proportion of M-phase cells and the subsequent loss of dividing cells under RNAi targeting *CDC20* (Fig. 4b, Extended Data Fig. 9a) or the anaphase-promoting complex/cyclosome (APC/C) subunit gene *CDC23*³⁰ indicate that APC/C inhibition remains rate limiting for progression from M-phase in planaria. The SAC-mediated regulation of *CDC20* in human cells involves the recruitment of MAD1 and MAD2 to the kinetochore by two molecular complexes thought to act in parallel, the broadly conserved KNL1–BUB3–BUB1 (KBB) complex and the ROD–ZW10–ZWILCH (RZZ) complex, which has been studied less because of its absence in yeast³¹ (Fig. 4a). The lack of clear *KNL1* and *MIS12* homologues, and of a cell-cycle phenotype of RNAi targeting *BUB3* (Fig. 4b), indicates that planarians have lost KBB complex function. However, we could identify clear RZZ complex homologues and, notably, knockdown of these

homologues prevented nocodazole-mediated M-phase arrest without affecting basal stem cell numbers or proliferation (Fig. 4b, Extended Data Fig. 9b). Therefore, planarian RZZ components control APC/C–*CDC20* either independently of MAD1 and MAD2 or in concert with homologues that have lost defining sequence features (Extended Data Figs 6, 7). Our results motivate the examination of putative MAD1 and MAD2-independent roles of the RZZ complex in other model systems and, together with the striking evolutionary plasticity of the SAC network in eukaryotes¹³, generally challenge our understanding of a core cell biological mechanism.

Discussion

We have described the highly contiguous genome sequence of the planarian model species *S. mediterranea*, which enables the genomic analysis of whole-body regeneration, stem cell pluripotency, lack of organismal ageing and other notable features of this model system. The resulting bird's eye view of a 'difficult' genome using long-read sequencing and *de novo* assembly also highlights important challenges that remain to be overcome. In the case of *S. mediterranea*, these include an abundance of low-complexity microsatellite repeats, inbreeding-resistant heterozygosity and a new class of extraordinarily long LTR elements. However, the fact that the scaffold size of newly reported genome assemblies often remains substantially below the 3.85 Mb of the *S. mediterranea* assembly (Extended Data Table 1) indicates that similar challenges may be widespread. We therefore expect that the specific improvements of the MARVEL assembler towards heterozygous and/or compositionally biased sequencing data¹⁵ will be useful for enhancing assembly contiguity in *de novo* genome sequencing projects.

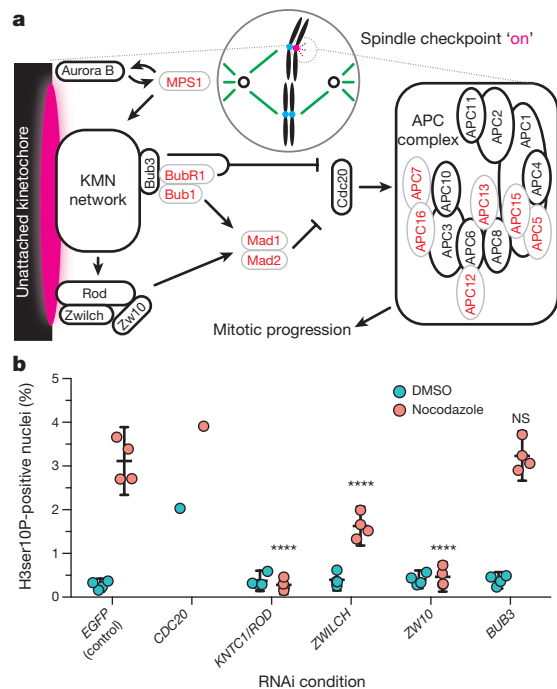


Figure 4 | Spindle assembly checkpoint (SAC) function in the likely absence of MAD1–MAD2. **a**, Cartoon illustration of SAC core components and function. Black and red denote components conserved or missing in *S. mediterranea*, respectively. KMN network: KNL1, MIS12 complex, NDC80 complex. **b**, Fractional abundance of mitotic cells under RNAi targeting the indicated SAC component genes, with (red) and without (cyan) nocodazole pre-treatment. Values are shown as mean with 95% confidence intervals ($n = 4$ biological replicates, 10 pooled animals, 5 technical replicates with 5 or 6 images each). Cells treated with RNAi targeting *CDC20* are shown as single replicates owing to rapid stem cell loss (Supplementary Information S18, Extended Data Fig. 9a, b). Significance assessed by two-way ANOVA, followed by Dunnett's post-hoc test (**** $P < 0.0001$; NS, not significant), excluding RNAi targeting *CDC20*.

We have also found a high degree of structural rearrangement and the absence of a number of conserved genes in the *S. mediterranea* genome. However, *D. melanogaster*, *C. elegans* and other animals also show loss of 'essential' genes^{13,26,32}, which raises a general conundrum: how can animals survive and compete while lacking core components of essential mechanisms? In cell biological terminology, a core mechanism signifies a chain of molecular interactions that explain a given process in multiple species, while essentiality indicates importance for organismal survival. The emergence of viable yeast strains upon deletion of essential genes³³ or the competitiveness of hundreds of extant planarian species in a diversity of habitats worldwide³⁴ both make it clear that essentiality is relative. The demonstration of SAC function in the likely absence of MAD1 and MAD2 suggests that our genetic and mechanistic understanding of SAC function is incomplete. Further studies on planarians and other 'non-traditional' model organisms are needed to understand the basis and mechanism of these cellular functions. Such a function-oriented, rather than gene-centric, view of biological mechanisms abstracts general function from individual molecules and is therefore likely to ultimately facilitate the reverse engineering of biology.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Data Availability The *S. mediterranea* genome assembly is accessible at GenBank under accession number NNSW000000000 and can also be browsed at and downloaded from <http://planmine.mpi-cbg.de>. All DNA and RNA reads were deposited at the Sequence Read Archive under the bioproject accession

PRJNA379262 and under the following SRA accession numbers: PacBio P4/C2 data, SRX2700681 and SRX2700682; PacBio P6/C4 data, SRX2700683; PacBio CCS data, SRX2700684; DNA shotgun, SRX2700686; DNA Chicago, SRX2700687; and RNA-seq, SRX2700685.

Code Availability The current source code of the MARVEL assembler is available at <https://github.com/schloi/MARVEL>. The execution scripts used for *S. mediterranea* can be found in the smed subfolder of the examples folder.

Received 6 April; accepted 21 December 2017.

Published online 24 January 2018.

- Rink, J. C. Stem cell systems and regeneration in planaria. *Dev. Genes Evol.* **223**, 67–84 (2013).
- Saló, E. & Agata, K. Planarian regeneration: a classic topic claiming new attention. *Int. J. Dev. Biol.* **56**, 3–4 (2012).
- Reddien, P. W. & Sánchez Alvarado, A. Fundamentals of planarian regeneration. *Annu. Rev. Cell Dev. Biol.* **20**, 725–757 (2004).
- Wagner, D. E., Wang, I. E. & Reddien, P. W. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science* **332**, 811–816 (2011).
- Onal, P. et al. Gene expression of pluripotency determinants is conserved between mammalian and planarian stem cells. *EMBO J.* **31**, 2755–2769 (2012).
- Tsai, I. J. et al. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57–63 (2013).
- Laumer, C. E., Hejnal, A. & Giribet, G. Nuclear genomic signals of the 'microturbellarian' roots of platyhelminth evolutionary innovation. *eLife* **4**, e05503 (2015).
- Collins, J. J., III & Newmark, P. A. It's no fluke: the planarian as a model for understanding schistosomes. *PLoS Pathog.* **9**, e1003396 (2013).
- Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
- Robb, S. M. C., Gotting, K., Ross, E. & Sánchez Alvarado, A. SmedGD 2.0: The *Schmidtea mediterranea* genome database. *Genesis* **53**, 535–546 (2015).
- Nishimura, O. et al. Unusually large number of mutations in asexually reproducing clonal planarian *Dugesia japonica*. *PLoS One* **10**, e0143525 (2015).
- Eid, J. et al. Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–138 (2009).
- van Hooff, J. J., Tromer, E., van Wijk, L. M., Snel, B. & Kops, G. J. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.* **18**, 1559–1571 (2017).
- Musacchio, A. & Salmon, E. D. The spindle-assembly checkpoint in space and time. *Nat. Rev. Mol. Cell Biol.* **8**, 379–393 (2007).
- Nowoshilow, S. et al. The axolotl genome and the evolution of key tissue formation regulators. *Nature* <https://doi.org/10.1038/nature25458> (2018).
- Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
- Putnam, N. H. et al. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
- Brandl, H. et al. PlanMine—a mineable resource of planarian biology and biodiversity. *Nucleic Acids Res.* **44**, D764–D773 (2016).
- Mouse Genome Sequencing Consortium Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
- Macas, J. & Neumann, P. Ogre elements—a distinct group of plant Ty3/gypsy-like retrotransposons. *Gene* **390**, 108–116 (2007).
- Guo, L., Zhang, S., Rubinstein, B., Ross, E. & Alvarado, A. S. Widespread maintenance of genome heterozygosity in *Schmidtea mediterranea*. *Nat. Ecol. Evol.* **1**, 0019 (2016).
- Wasik, K. et al. Genome and transcriptome of the regeneration-competent flatworm, *Macrostomum lignano*. *Proc. Natl Acad. Sci. USA* **112**, 12462–12467 (2015).
- Lai, A. G., Kosaka, N., Abnave, P., Sahu, S. & Aboobaker, A. A. The abrogation of condensin function provides independent evidence for defining the self-renewing population of pluripotent stem cells. *Dev. Biol.* **433**, 218–226 (2018).
- Ceccaldi, R., Rondinelli, B. & D'Andrea, A. D. Repair pathway choices and consequences at the double-strand break. *Trends Cell Biol.* **26**, 52–64 (2016).
- Stubenhaus, B. M. et al. Light-induced depigmentation in planarians models the pathophysiology of acute porphyrias. *eLife* **5**, e14175 (2016).
- Rao, A. U., Carta, L. K., Lesuisse, E. & Hamza, I. Lack of heme synthesis in a free-living eukaryote. *Proc. Natl Acad. Sci. USA* **102**, 4270–4275 (2005).
- Grudniewska, M. et al. Transcriptional signatures of somatic neoblasts and germline cells in *Macrostomum lignano*. *eLife* **5**, e20607 (2016).
- Santaguida, S. & Amon, A. Short- and long-term effects of chromosome mis-segregation and aneuploidy. *Nat. Rev. Mol. Cell Biol.* **16**, 473–485 (2015).
- McWhinnie, M. A. & Gleason, M. M. Histological changes in regenerating pieces of *Dugesia dorotocephala* treated with colchicine. *Biol. Bull.* **112**, 371–376 (1957).
- Kang, H. & Sánchez Alvarado, A. Flow cytometry methods for the study of cell-cycle parameters of planarian stem cells. *Dev. Dyn.* **238**, 1111–1117 (2009).

31. Silió, V., McAinsh, A. D. & Millar, J. B. KNL1-Bubs and RZZ provide two separable pathways for checkpoint activation at human kinetochores. *Dev. Cell* **35**, 600–613 (2015).
32. Sekelsky, J. DNA repair in *Drosophila*: mutagens, models, and missing genes. *Genetics* **205**, 471–490 (2017).
33. Rancati, G. *et al.* Aneuploidy underlies rapid adaptive evolution of yeast cells deprived of a conserved cytokinesis motor. *Cell* **135**, 879–893 (2008).
34. Schockaert, E. R. *et al.* Global diversity of free living flatworms (Platyhelminthes, 'Turbellaria') in freshwater. *Hydrobiologia* **595**, 41–48 (2008).
35. Wurtzel, O. *et al.* A generic and cell-type-specific wound response precedes regeneration in planarians. *Dev. Cell* **35**, 632–645 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J.-H. Lee for multiple sequence alignments; T. Boothe and M. V. Farré for karyotyping; A. Hejnol, A. Desai and J. Mansfeld for critical reading of the manuscript; and DoveTail Genomics staff for graphical support. We thank the following MPI-CBG facilities for their support: DNA sequencing, Scientific computing and Light microscopy. We thank V. Benes and the EMBL GeneCore and A. Dahl and the Deep Sequencing Group (SFB 655/BIOTEC) for RNA sequencing, and S. von Kannen, H. Andreas, S. Clausen and N. Gscheidel for technical support. This project received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement number 649024) and the Max Planck Society. G.R.Y. was supported by the Francis Crick Institute under

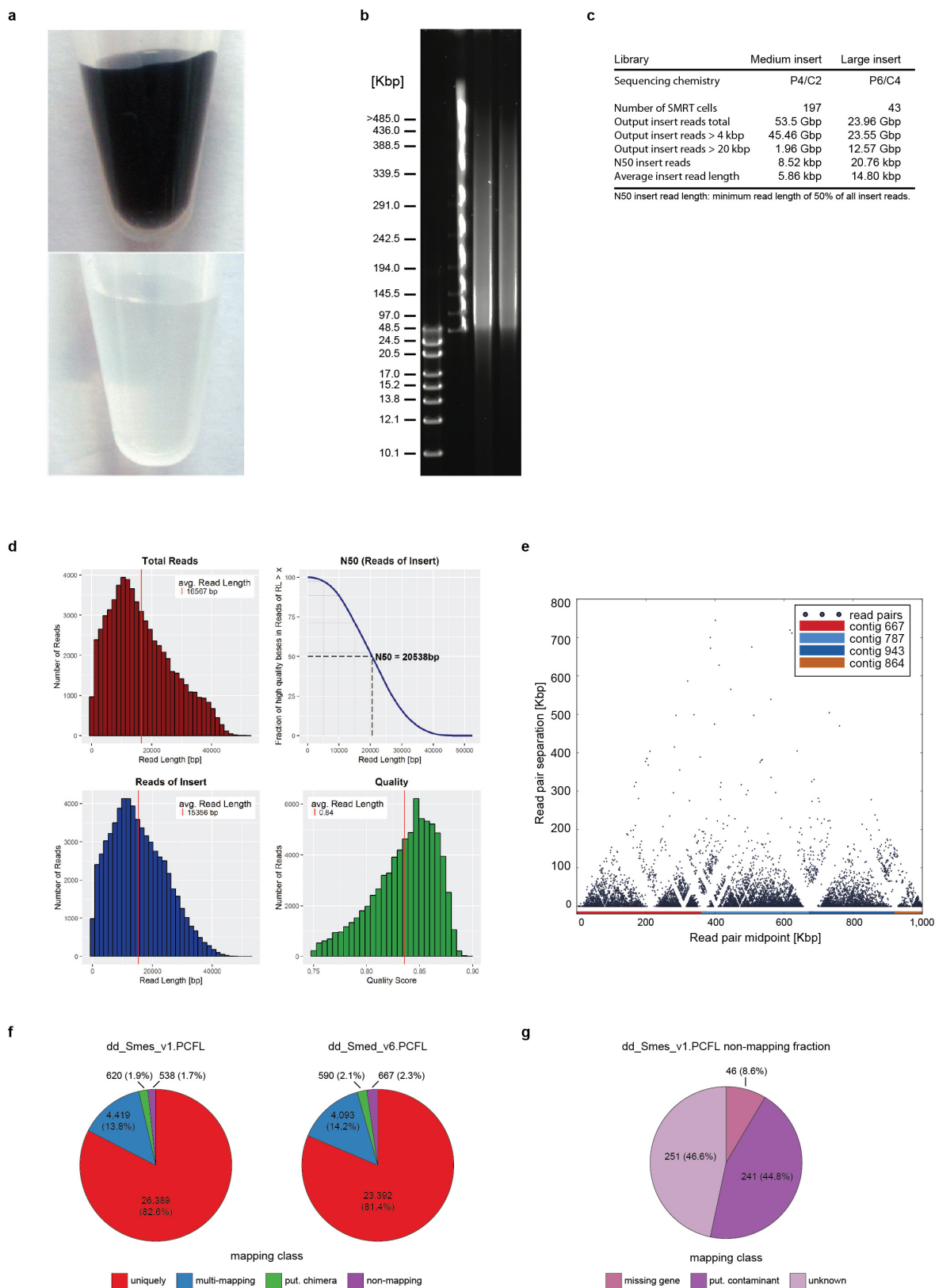
award FC001162 (J. P. Stoye). The Crick Institute receives its core funding from Cancer Research UK, the UK Medical Research Council and the Wellcome Trust.

Author Contributions Conceptualization: J.C.R., E.M., S.S.; methodology: M.A.G., M.P., A.R., G.R.Y., S.W., H.B., I.H., M.H., J.C.R.; formal analysis: M.A.G., M.P., A.R., G.R.Y., H.B., I.H., S.W., M.H.; investigation: M.A.G., M.P., A.R., M.H., J.C.R.; writing (original draft): J.C.R.; writing (review and editing): J.C.R., E.M., M.H., S.P.; visualization: J.C.R., M.H., S.P., I.H., H.B., S.W., G.R.Y., A.R., M.P., M.A.G.; funding acquisition: J.C.R., E.M., S.S., A.D. All authors read and approved the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.C.R. (rink@mpi-cbg.de), E.M. (myers@mpi-cbg.de) or S.S. (siegfried.schloissnig@h-its.org).



This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons licence, users will need to obtain permission from the licence holder to reproduce the material. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

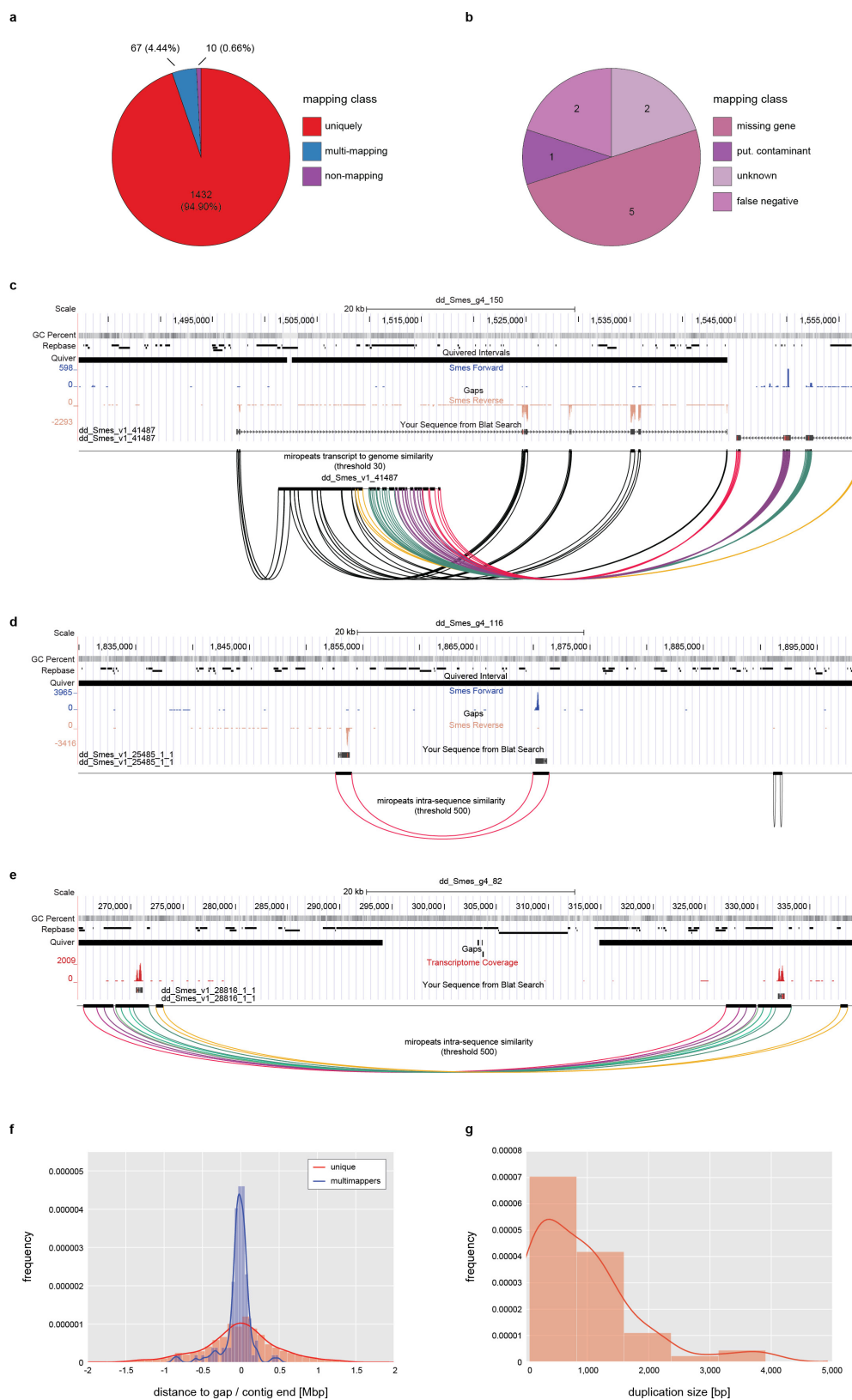


Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | *S. mediterranea* sequencing and assembly quality control.

a. *S. mediterranea* genomic DNA preparation. The established protocol (top) yields a black solution owing to co-purification of porphyrin pigments. The improved protocol (bottom) removes contaminants including the pigment and therefore results in clear preparations. **b.** The improved protocol consistently yields high-molecular-weight DNA, as shown by the pulse field gel electrophoresis of two independent preparations (right-hand two lanes) and DNA size markers (left-hand two lanes). **c.** Overview of all PacBio sequencing runs for the *S. mediterranea* assembly. **d.** Sequencing statistics of a representative PacBio RS II SMRT cell (P6/C4 chemistry). Total output: 1,053.4 Mb; reads of insert: 976.4 Mb; maximal read length: 52,441 bp. **e.** Connectivity matrix plot illustrating Chicago library read-pair distances after HiRise scaffolding. Colour coding identifies individual contigs contributing to the scaffold dd_Smed_g4_1. **f.** Mapping characteristics of *S. mediterranea* transcriptomes against the genome assembly with more than 60% query coverage and more than 60% sequence identity as cut-off criteria. Left, the dd_Smes_v1.PCFL transcriptome of the

sequenced strain. Right, the dd_Smed_v6.PCFL transcriptome of the asexual strain. The pie charts show the absolute number and relative proportions of transcripts mapping with the indicated characteristics. **g.** Further analysis of the 538 non-mapping *S. mediterranea* transcripts from **e** (Supplementary Information S7). Missing gene, transcripts that map uniquely to the SmedSxl v4.0 assembly¹⁰ and have annotated orthologues in at least five other planarian species in PlanMine¹⁸. Putative contaminant, Top RefSeq BLAST hit in a likely contaminant species. Unknown, all remaining transcripts. The fact that only 46 out of 31,966 *S. mediterranea* transcripts are classified as genuinely missing indicates that the *S. mediterranea* assembly is largely complete. In contrast, 1,229 transcripts that uniquely mapped to the *S. mediterranea* genome and had orthologues in at least five other planarian species failed to map to the previously published SmedSxl v4.0 assembly¹⁰. Substantial gaps in the previous assembly also mean that the number of missing genes in the *S. mediterranea* assembly may be slightly higher, as some may have been classified as unknown.

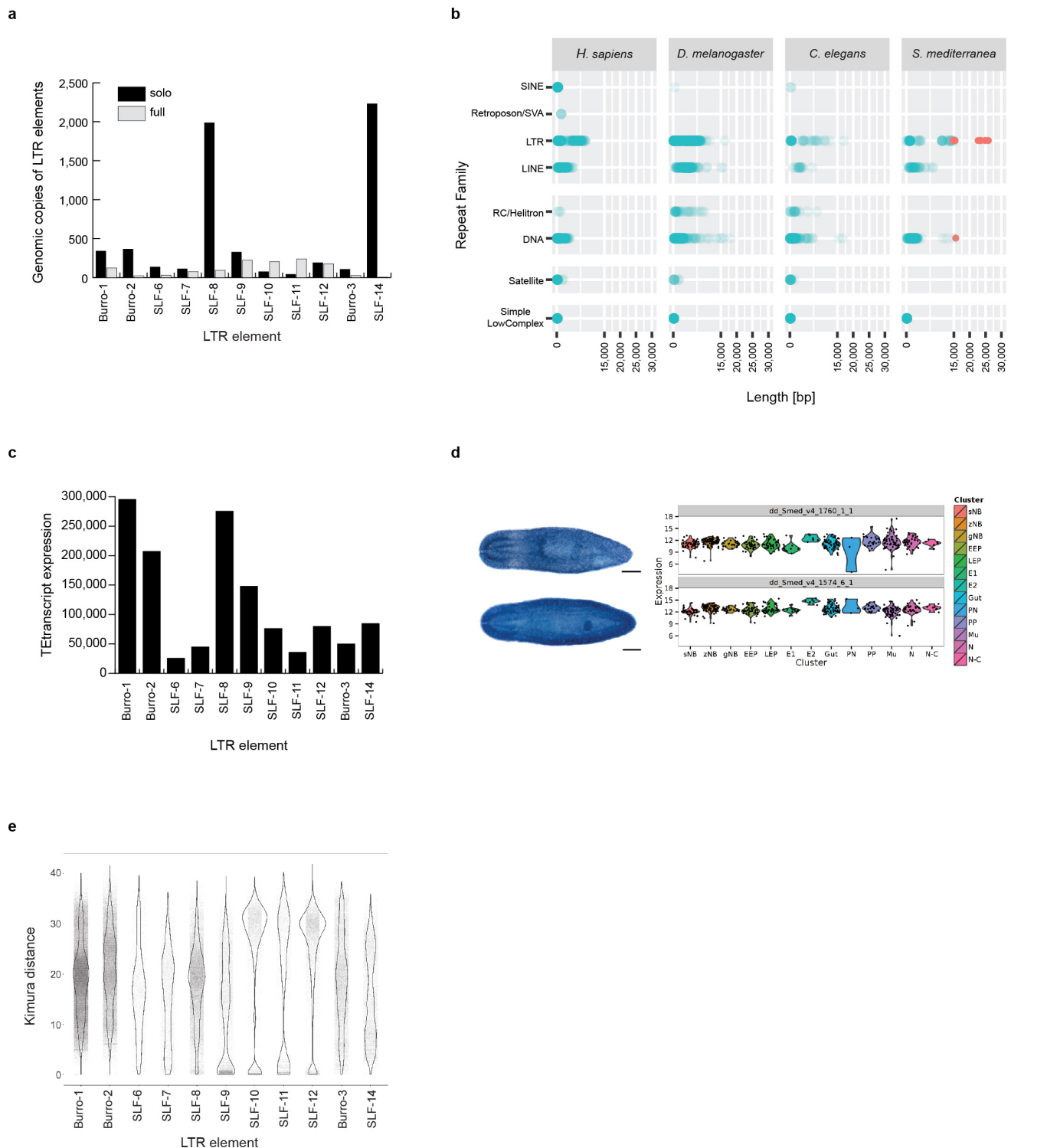


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Assembly validation by high stringency transcript back-mapping.

a, Quality control of the *S. mediterranea* assembly by means of high stringency back-mapping of 1,509 high confidence (HC)-cDNAs. HC-cDNAs were defined as having BLAST hits with more than 90% query and subject coverage in seven other planarian transcriptomes in PlanMine¹⁸. HC-cDNAs were mapped to the *S. mediterranea* assembly using more than 90% query coverage and sequence identity as cut-off criteria. The pie chart shows the absolute number and relative proportions of HC-cDNAs mapping with the indicated characteristics. **b**, Further analysis of the ten HC-cDNAs classified as non-mapping from **a** by intersection with the mapping results of Extended Data Fig. 1g. Of these, two were designated as 'false negative' as both mapped to the *S. mediterranea* genome with more than 90% query coverage and sequence identity using BLAT. **c**, UCSC genome browser screenshot (75-kb window) of the genomic mapping location of one of the two 'unknown' HC-cDNAs as a single example of a mapping failure due to an actual assembly error. The example documents inversion

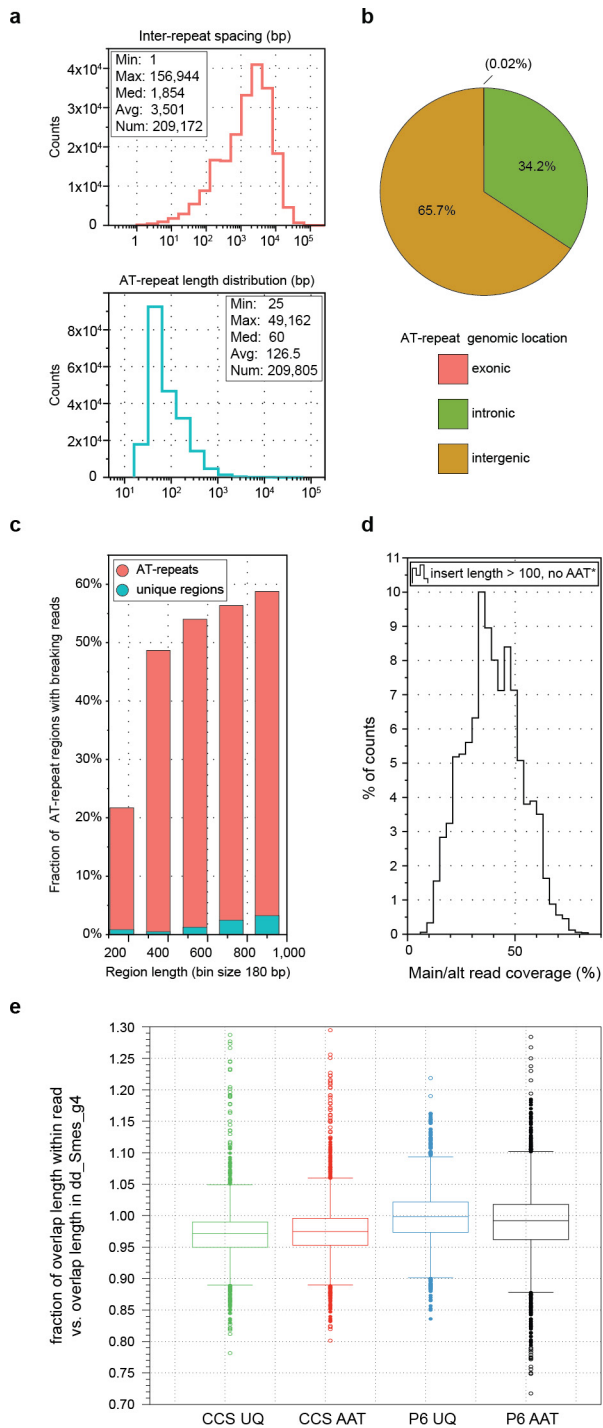
of the 5'-end of the cDNA within a low-confidence stretch at a contig end (lack of coverage in the Quiver track). The inversion is supported by inverted RNA-seq read mapping and inversion of the cDNA sequence shown in the respective tracks. Below, colour-coded Miropeats similarity plots of respective regions. **d**, **e**, Examples of genomic mapping loci of HC-cDNA transcripts from the multi-mapping category in **a**, browser screen shots as described in **c**. **d**, Example of a likely legitimate (biological) gene duplication in a gap-free high-confidence region. **e**, Micro tandem duplication surrounding a scaffolding gap in a repeat-rich region. **f**, Multi-mapping HC-cDNAs map preferentially to contig ends. The histogram plots the distance to the closest gap or contig end for the 67 multi-mappers and a corresponding number of unique mappers (**a**). **g**, Estimated size of the duplicated regions of multi-mapping HC-cDNAs. This analysis identifies a small fraction of small-scale duplications at assembly gaps in the *S. mediterranea* assembly, which can be easily identified with the help of the various quality control tracks in the PlanMine genome browser.



Extended Data Figure 3 | Repeats in the *S. mediterranea* assembly.

a, Estimation of the abundance of solo and full-length LTR elements in the *S. mediterranea* assembly. Elements SLF-8 and SLF-14 show a large number of solo LTRs compared to full-length copies, indicating a large number of excision events by homologous recombination. Of the Burro elements, Burro-1 was the most abundant, with 124 full-length copies, followed by Burro-3 and Burro-2 with 25 and 23 full-length copies, respectively. **b**, Comparison of lengths of indicated repeat consensus classes in *H. sapiens*, *D. melanogaster* and *C. elegans*. For *S. mediterranea*, we used a custom library generated in this study. Dark colours indicate predominant lengths of specific repeat classes. Red, repeat consensus more than 15 kb in length. **c**, Expression analysis of gypsy LTR elements in *S. mediterranea* RNA-seq data using TETranscripts. The three most transcriptionally active elements were Burro-1, Burro-2 and SLF-8.

d, LTR expression analysis by whole-mount *in situ* hybridization and single-cell expression data³⁵. Top, SLF-9-derived transcript. Bottom, Burro-1-derived transcript. Both are broadly transcribed in many *S. mediterranea* cell types (CIW4 strain, $n = 1$ biological replicate, 10 animals). Scale bar, 250 μm . **e**, Kimura distance plot of *S. mediterranea* LTR elements. Substitution levels varied by element, but also within element groups. Burro-1, Burro-2, Burro-3 and SLF-8 all contain elements spread over a large range of substitution levels, possibly indicative of continued activity over large time scales. The remaining elements are characterized by more defined peaks in expansion, with the highest average divergences being seen in the smallest elements characterized (SLF-10, SLF-11, SLF-12), making these amongst the oldest within the genome. Notably, both SLF-8 and SLF-9 have representative elements with particularly low substitution rates, potentially indicating a recent or ongoing expansion.

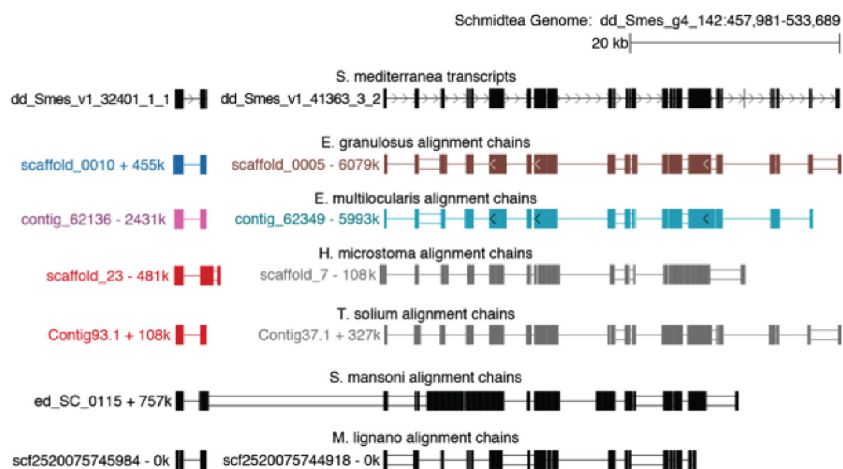


Extended Data Figure 4 | AT-rich microsatellites in the *S. mediterranea* genome. **a**, Features of AT-rich microsatellites. Top, inter-repeat spacing of repeats more than 99 bp in length. Bottom, repeat length. AT-rich microsatellites with an average length of 127 bp occur every approximately 3,500 bp. **b**, Genomic distribution of repeats more than 99 bp in length. **c**, Increased probability of read alignment termination within microsatellite repeats. Individual size bins were analysed separately for microsatellite repeats (red) or non-repetitive regions (cyan). Although accounting for only 4.2% of the assembly size, microsatellite repeats significantly limit assembly contiguity owing to an increased probability of read alignment loss. **d**, Genome-wide coverage ratios of insertion or deletion sequences more than 99 bp in length and excluding AT repeats. **e**, Read length variation analysis across AT-rich repeat regions (AAT) in regular PacBio sequencing data compared to circular consensus sequencing (CCS) coverage of the same region. CCS reads sample the same genomic region multiple times. The lack of a clear difference in the length variation of specific AT repeats between repetitive sequencing of the same DNA molecule (CCS) and that of sequencing reads representing different DNA molecules (regular PacBio data, P6/C4) indicates that repeat length variations are mainly technical in nature. Rather than repeat length polymorphisms, the most likely cause of the detrimental effect of the repeats is the increased ambiguity in low complexity sequence alignments (Supplementary Information S11.4). Unique (UQ) regions were included as controls. CCS_UQ: CCS subread length variation versus the consensus length of all subreads in binned unique regions ($n = 3,300$). CCS_AAT: CCS subread length variation versus the consensus length of all subreads in binned AT repeat regions ($n = 4,825$). P6_UQ: Length variation of individual reads in the regular PacBio sequencing data (P6/C4) versus the consensus length of the region in the *S. mediterranea* assembly in binned unique regions ($n = 3,310$). P6_AAT: Length variation of individual reads in the regular PacBio sequencing data (P6/C4) versus the consensus length of the region in the *S. mediterranea* assembly in binned AT repeat regions ($n = 5,085$). Dots, outliers; central horizontal line, median; box ranges, from first quartile to third quartile; whiskers, interquartile range (midspread): 75th and 25th percentile.

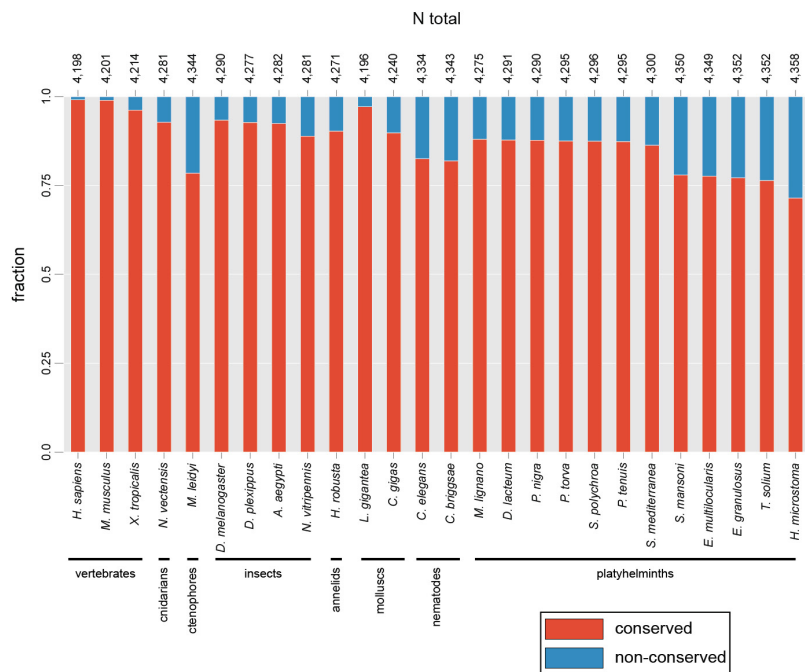
a

	contig stats		scaffold stats	
	number	N50 [kbp]	number	N50 [kbp]
<i>E. granulosus</i>	1,288	5,229	1,288	5,229
<i>E. multilocularis</i>	1,855	4,914	1,815	4,914
<i>H. microstoma</i>	1,132	539	1,132	539
<i>T. solium</i>	5,508	182	5,508	182
<i>S. japonica</i>	93,729	6	25,048	177
<i>S. mansoni</i>	9,517	77	881	32,115
<i>S. mediterranea</i>	1,292	1,122	481	3,855
<i>M. lignano</i>	36,448	4	17,663	26
<i>G. gallus (chicken)</i>	29,829	273	15,931	90,217
<i>X. tropicalis (frog)</i>	55,149	72	7,727	124,127
<i>D. rerio (zebrafish)</i>	22,353	1,324	1,060	53,345
<i>P. marinus (lamprey)</i>	73,813	13	25,005	185

b



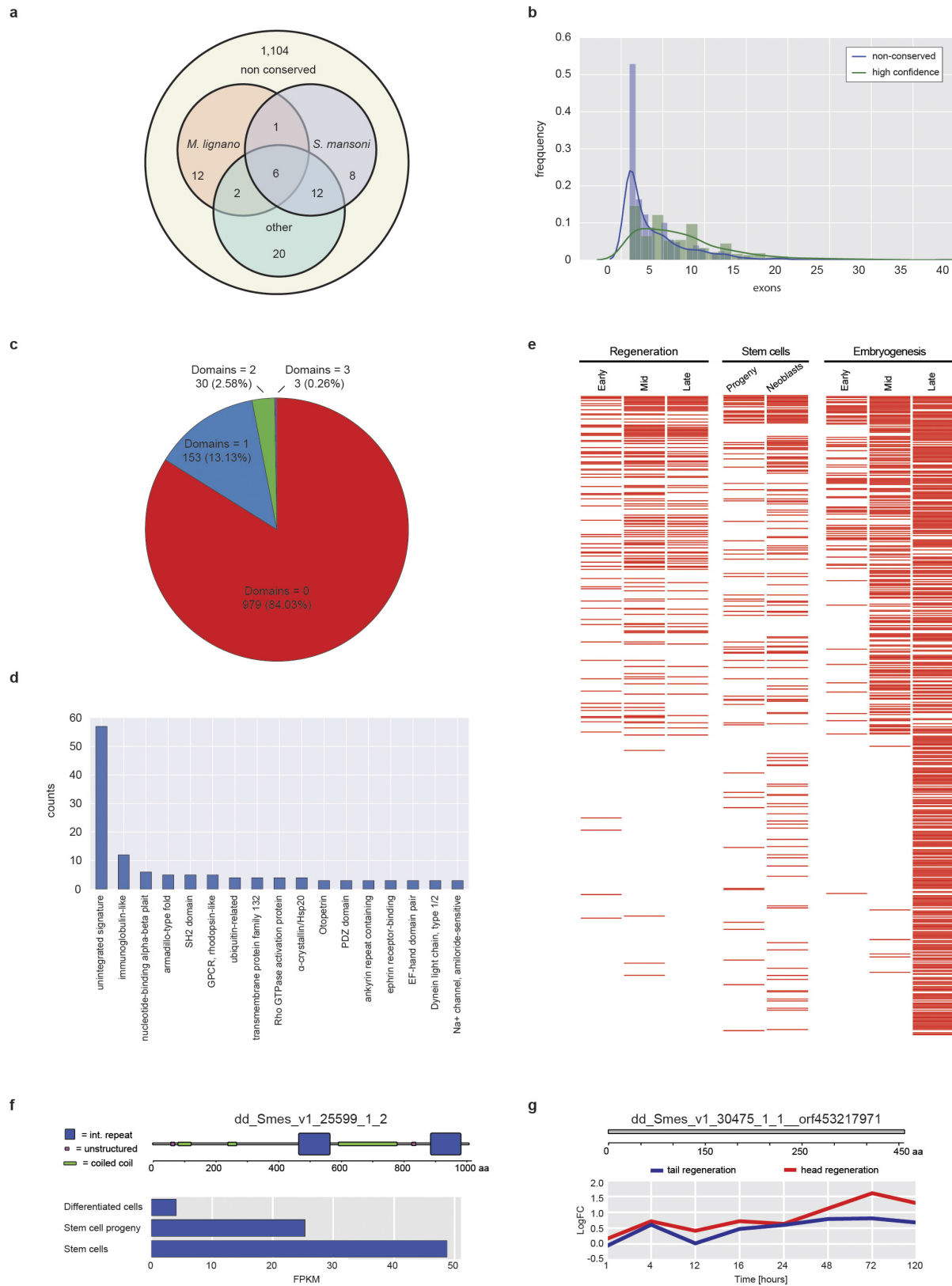
c



Extended Data Figure 5 | See next page for caption.

Extended Data Figure 5 | Comparative genomics. a, Contig and scaffold N50 statistics of the genomes used for the comparative genome alignments in Fig. 3b. The basal vertebrate lamprey genome assembly is more fragmented (similar or lower N50 values) than most other platyhelminth genomes. Nevertheless, the human-to-lamprey genome alignment has equivalent or even higher alignment chain scores and span lengths than any of the platyhelminth genome alignment comparisons, indicating that the true extent of sequence divergence and loss of conserved gene order in platyhelminths is likely to be greater than estimated. **b,** Example of a top-scoring alignment chain. The UCSC genome browser screenshot of the *S. mediterranea* genome shows that alignments predominantly overlap exons of the two transcripts shown at the top. This example is one of the few cases of apparent gene order conservation between *S. mediterranea*

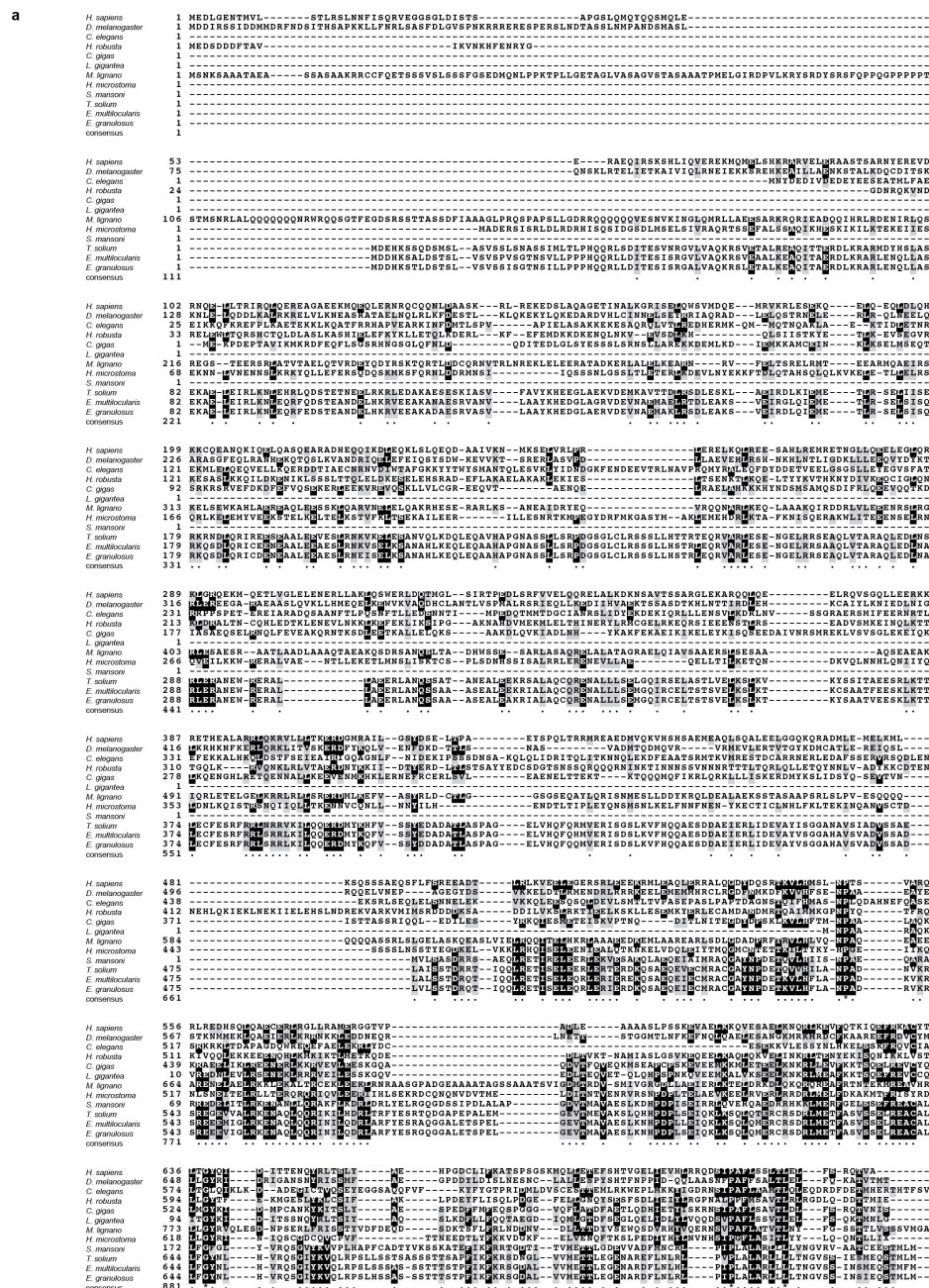
and *S. mansoni*. Blocks in the alignment chains represent local alignments, connecting single lines represent deletions in the query genome and double lines represent regions with sequence in both *S. mediterranea* and the query genome that do not align. **c,** Comparative loss analysis of highly conserved genes across the 26 indicated species. Red, conserved gene fraction, defined as the proportion of orthogroups containing at least 9 out of the 14 non-flatworm species and the query species. Blue, lost fraction of highly conserved genes, defined as the proportion of orthogroups containing at least 9 out of the 14 non-flatworm species, but not the query species (Supplementary Information S17). Absolute numbers of highly conserved genes are shown at the top, with slight fluctuations caused by species-specific sequence duplications.



Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Planarian-specific genes. **a**, Conservation of 1,165 flatworm-specific genes (Supplementary Information S16.1) amongst flatworm species. Only 61 sequences had sequence homologues in the indicated flatworm species (other denotes *Taenia solium*, *Echinococcus multilocularis*, *Echinococcus granulosus*, *Hymenolepis microstoma*), indicating that this gene set mostly represents planarian-specific genes. **b**, **c**, Characteristics of planarian-specific genes. **b**, Distribution of exon numbers compared to a control gene set (HC-cDNAs; Extended Data Fig. 2a), indicating enrichment of single-exon genes. **c**, Number of predicted domains (InterProScan), indicating that only a minority of genes contain predicted domains. **d**, Identity of detected domains (Pfam and SUPERFAMILY). Unintegrated signatures,

recurring sequence motifs that are not grouped into InterPro entries. These might represent so far uncured or weakly supported motifs that do not pass InterPro's integration standards. **e**, Differential expression of 626 planarian-specific genes in published *S. mediterranea* RNA-seq data sets of different regeneration phases (left), stem cells or progeny populations (middle) or specific developmental stages (right). Red lines indicate differential expression relative to the control of each series (white indicates no change). Genes were ordered using rank by sum. The high proportion of differential expression indicates the widespread contribution of lineage-specific genes to planarian biology. **f**, **g**, Specific examples of non-conserved genes. Top, SMART domain representation. Bottom, Differential expression under the indicated conditions.



Extended Data Figure 7 | Sequence conservation of MAD1 protein in non-planarian flatworms. a. COBALT multiple protein sequence alignment of the MAD1 homologues of the indicated species (including all the non-planarian flatworm species from Fig. 3c). **b.** Heat map of

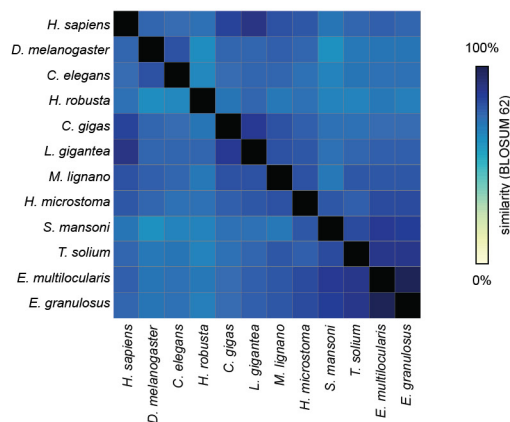
BLOSUM62 sequence similarity matrix generated from alignment in **a**, demonstrating significant sequence conservation of MAD1 homologues even in flatworms.

a

H. sapiens 1 MAL-QLSREOGITLRCGSAITVAEFLPSFGINSILYORGIYFSETTRVQVYGLTLVTTDLLEIKYINNVEQLKD-----WLYKCSVOKLVVVISNIESGGEVLERWQFD
D. melanogaster 1 MST-AQAHKNCIDLKGSATIIVEYLKYGINSILYORGIYPAEDDNTQVYGLTLHSKDPKIKTFLONVLSQTEE-----WLSKNMINKISMVITNAHKEVLECWQFD
C. elegans 1 MTD-VK--HONATSLKGSALVKEFFHFGINSILYORGIYPAEDDNTQVYGLTLVVAHEKKKQAFMDPLQOQVEE-----WIAKROIKRLVMVISEVKKKEVVERWQFD
H. robusta 1 -----MYPEDMPTRVVDFGVPLMVVVKESDYTDQHSKYTEE-----WIEKLTQKRLVMTSRVDTGDIWERWQFD
C. gigas 1 MA--ATAHSNATLTKGSTIVASFLIGINSILYORGIYFSETTRVQVYGLTLVTTSDPKKDYENPVIAGTKE-----WLYNMAVKKLVVIRKVDNEVLERWQFD
L. gigantea 1 MA--QLKRNATLTKGSTIVVDFHFGINSILYORGIYFSETTRVQVYGLTLVTTSDCKEIDNITSQKE-----WLTNMSQKRLVVVKSHTNEVLERWQFD
M. lignano 1 MPV-AQAKHATLTKGSTIVSEFFIYVNSILYORGIYFSEHSFEQVYGLTLVTTSDENKKAVIKILEQVKQ-----WISDLVTEKLVVIRVSGGEVVERWQFD
S. mansoni 1 MPT---EASATSLKGSALLTDYFFIYVNSILYORGIYFSAFQKNIKYDLSVLVTDENLIKYLNVILNOVKN-----WLEDGSVHRLALIKSVKSGEVLERWQFD
H. microstoma 1 -----WIMSDSLKRLVVIVIKSVKSGEATLERWQFD
T. solium 1 --MQAPQLSNATLTKGSVSVIGDYFQVAINNLLIRGIYFSEASFKOVKKNFERSVLTHTDELDIDYDLSLSQVKASLSLLVWISGGSLKRLVIVIKSVKTECHLERWQFD
E. multilocularis 1 MALQKLQNSNATLTKGSVEVIADYFYVAINNLLIRGIYFSEASFKOVKKNFERSVLTHTDELDIDFNCLVROMK-----WVWSSDSLKRLVIVIKSAKTEVLERWQFD
E. granulosus 1 MALQKLQNSNATLTKGSVEVIADYFYVAINNLLIRGIYFSEASFKOVKKNFERSVLTHTDELDIDFNCLVROMK-----WVWSSDSLKRLVIVIKSAKTEVLERWQFD
consensus 1

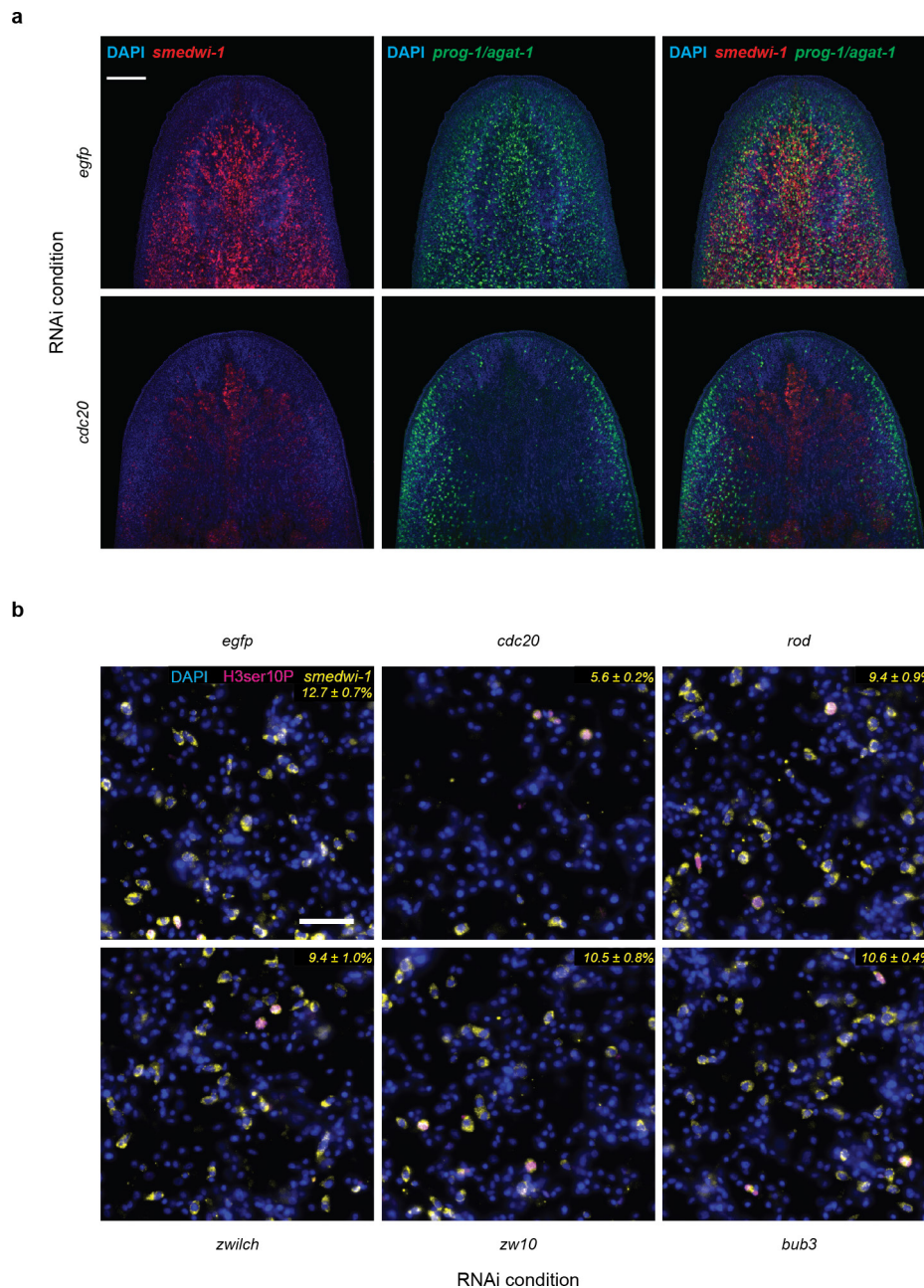
H. sapiens 104 IECDDTAKDD-----SAPREHSKAMODEPFRSVIRQITATVTFPLLEVSQSDLLITVTKDLVVP--HKMDESGPOLITNSEEVRLSFTTHHKVNSMVAKTIPVND
D. melanogaster 104 MQAELGDDDIS--DPTKATTTKELSRITONEIRDVIRQISATVSYLPLDLCICTFDDNHHFLONTSLP--AKMDETGAVIONPQAVOLRSFSTGGHKVDLVVNYKMS--
C. elegans 103 IHTB---NLAE--EGENAHRTKEKKRQCEISDVIRQISATVSYLPLDEEPVSPDVHIVYCKQTQAP--EDWTEGACLIQNETVOLRSFSTSVHSVNTNVOYKADF--
H. robusta 68 VECDDTATHHTI-----KESTEADVRKSDQVIRQISATVSYLPLDPCPSSEFLHRTKSSQVAP--EYKKEAGLQNSLSEKRWSTTHHKVSTSVAKKLD--
C. gigas 103 VEGDTEALKDGYLNDCKSPROSEKDNNEISAVIRQISATVSYLPLDEGACADPLVYTDKDLDP--AKMGGDUTQPIANSSEVRLSFTTHHKVDAMVAFKK--
L. gigantea 103 IDCDDTATVSE-----TKRMDEKEKDKDKSVIRQISATVSYLPLDETACAFDPLVYTDKDLDP--ISWGGSGPHFVNSSEVRLSFTTHHKVDAMVAKKCD--
M. lignano 104 IKCDHSKSDSEF-----MVAEIDEKQKKDIADVIRQISATVSYLPLDETACAFDPLVYTDKDLDP--ISWGGSGPHFVNSSEVRLSFTTHHKVDAMVAKKCD--
S. mansoni 102 MTTEKTDSSNSV-----GTSLAQMSEIQGVIRQISATVSYLPLDSSCTFELLVIADRNANVP--TSWDETGPOIVNSDEKLRSSSTTHHKVDHIVSYKRS--
H. microstoma 30 IIRBEETAVGD-----HDEGLGALVROIVASSTPLNIEKRCCTFDLLVY--NRNCEIP--EGWDESGPQVPSACVOLRSFSTTHHKVSEIVSYRQK--
T. solium 109 VVRBEESFSRP-----HSVEQVNNHGLDIROIVSSSTPLNIEKRCCTFDLLVY--NRNCEIP--EGWDESGPQVPSACVOLRSFSTTHHKVSEIVSYRQK--
E. multilocularis 105 VNMDEESSVEP-----HNMDOVNNELSVIRQISATVSYLPLDSSCTFELLVIADRNANVP--TSWDETGPOIVNSDEKLRSSSTTHHKVSEIVSYRQK--
E. granulosus 105 VVKCEESSVEP-----HNMDOVNNELSVIRQISATVSYLPLDSSCTFELLVIADRNANVP--TSWDETGPOIVNSDEKLRSSSTTHHKVSEIVSYRQK--
consensus 111

b



Extended Data Figure 8 | Sequence conservation of MAD2 protein in non-planarian flatworms. a. COBALT multiple protein sequence alignment of the MAD2 homologues of the indicated species (including all the non-planarian flatworm species from Fig. 3c). **b.** Heat map of

BLOSUM62 sequence similarity matrix generated from alignment in a), demonstrating significant sequence conservation of MAD2 homologues even in flatworms.



Extended Data Figure 9 | Effect of RNAi targeting *CDC20* and SAC components on the planarian stem cell compartment. **a**, Fluorescent whole-mount *in situ* hybridization of the planarian head region. Stem cells (neoblasts) were visualized using a *smedwi-1* probe (red), early and late progeny by pooled *prog-1* and *agat-1* probes (green). Blue, nuclear counterstaining by DAPI. Top, RNAi control against *EGFP*; bottom, RNAi targeting *CDC20*, which results in a markedly decreased number of *smedwi-1* and *prog-1/agat-1* positive cells after three rounds of RNAi feeding. This indicates the loss of neoblasts and a concomitant reduction

in progenitor numbers ($n = 1$ biological replicate, 10 animals). Scale bar, 200 μm . **b**, Effect of indicated RNAi treatments on planarian stem cell abundance. Representative images of cell macerates, stained with DAPI (nuclei, blue), anti-H3ser10P (mitotic cells, magenta) and *smedwi-1* *in situ* hybridization (stem cells, yellow). Numbers indicate the mean fraction \pm s.d. of *smedwi-1* positive cells of total cells quantified by nuclear counting using DAPI ($n = 1$, 10 pooled animals, 5 technical replicates with 5 images each). Scale bar, 50 μm .

Extended Data Table 1 | *S. mediterranea* genome assembly comparisons

Assembly	SmedSx1 v4.0	GCA_000691995.1	PacBio-Canu	PacBio - MARVEL	g4 assembly
Technology	Sanger	Illumina	PacBio	PacBio	PacBio + Chicago
Assembler	NA	SOAPdenovo	Canu	MARVEL	MARVEL + HiRise
Assembly length (Mb)	787.5	700.7	938.8	782.1	774.0
Contigs					
# contigs	112,641	108,794	7,637	1,839	1,292
Longest contig	149,108	132,070	2,212,985	4,363,926	5,343,607
Contig N50	11,977	10,721	194,023	708,691	1,121,568
Scaffolds					
# scaffolds	15,334	12,782	NA	NA	481
Longest scaffold	893,023	1,050,243	NA	NA	17,761,579
Scaffold N50	80,447	83,932	NA	NA	3,854,845
% in gaps	13.87	14.32	0	0	0.01

Final Smed_g4 assembly characteristics are shown in bold.

Evolutionary routes and *KRAS* dosage define pancreatic cancer phenotypes

Sebastian Mueller^{1,2*}, Thomas Engleitner^{1,2,3*}, Roman Maresch^{1,2,3*}, Magdalena Zukowska^{1,2}, Sebastian Lange^{1,2}, Thorsten Kaltenbacher^{1,2,3}, Björn Konukiewicz⁴, Rupert Öllinger^{1,2}, Maximilian Zwiebel², Alex Strong⁵, Hsi-Yu Yen^{3,6}, Ruby Banerjee⁵, Sandra Louzada⁵, Beiyuan Fu⁵, Barbara Seidler^{1,2}, Juliana Götzfried², Kathleen Schuck², Zonera Hassan², Andreas Arbeiter², Nina Schönhuber^{1,2}, Sabine Klein^{1,2}, Christian Veltkamp^{1,2}, Mathias Friedrich⁵, Lena Rad², Maxim Barenboim^{2,3}, Christoph Ziegenhain⁷, Julia Hess⁸, Oliver M. Dovey⁵, Stefan Eser², Swati Parekh⁷, Fernando Constantino-Casas⁹, Jorge de la Rosa^{5,10,11}, Marta I. Sierra¹², Mario Fraga^{12,13}, Julia Mayerle¹⁴, Günter Klöppel⁴, Juan Cadiñanos^{5,10}, Pentao Liu⁵, George Vassiliou⁵, Wilko Weichert^{3,4}, Katja Steiger^{4,6}, Wolfgang Enard⁷, Roland M. Schmid^{2,3}, Fengtang Yang⁵, Kristian Unger⁸, Günter Schneider^{2,3}, Ignacio Varela¹⁵, Allan Bradley⁵, Dieter Saur^{1,2,3§} & Roland Rad^{1,2,3§}

The poor correlation of mutational landscapes with phenotypes limits our understanding of the pathogenesis and metastasis of pancreatic ductal adenocarcinoma (PDAC). Here we show that oncogenic dosage-variation has a critical role in PDAC biology and phenotypic diversification. We find an increase in gene dosage of mutant *KRAS* in human PDAC precursors, which drives both early tumorigenesis and metastasis and thus rationalizes early PDAC dissemination. To overcome the limitations posed to gene dosage studies by the stromal richness of PDAC, we have developed large cell culture resources of metastatic mouse PDAC. Integration of cell culture genomes, transcriptomes and tumour phenotypes with functional studies and human data reveals additional widespread effects of oncogenic dosage variation on cell morphology and plasticity, histopathology and clinical outcome, with the highest *Kras*^{MUT} levels underlying aggressive undifferentiated phenotypes. We also identify alternative oncogenic gains (*Myc*, *Yap1* or *Nfkb2*), which collaborate with heterozygous *Kras*^{MUT} in driving tumorigenesis, but have lower metastatic potential. Mechanistically, different oncogenic gains and dosages evolve along distinct evolutionary routes, licensed by defined allelic states and/or combinations of hallmark tumour suppressor alterations (*Cdkn2a*, *Trp53*, *Tgfβ*-pathway). Thus, evolutionary constraints and contingencies direct oncogenic dosage gain and variation along defined routes to drive the early progression of PDAC and shape its downstream biology. Our study uncovers universal principles of *Ras*-driven oncogenesis that have potential relevance beyond pancreatic cancer.

PDAC is the fourth leading cause of cancer-related death and is expected to become the second within the next decade¹. While treatments have constantly improved for many other cancer types, 5-year survival rates in PDAC have stayed around 5% (ref. 1). Genome sequencing revealed extensive genetic heterogeneity beyond a few frequently mutated drivers^{2–8} such as *KRAS*, *TP53*, *CDKN2A* or transforming growth factor- β (TGF β)-pathway alterations. Disappointingly, however, genomic changes cannot so far be broadly linked to biological, morphological or clinical phenotypes. In addition, the molecular basis of cancer cell dissemination is poorly understood, and genetic comparisons of primary–metastasis pairs could not identify recurrent alterations linked to metastasis^{3,8}. Critical limitations to human PDAC (hPDAC) genomics are (1) the complexity of cancer genomes, which poses challenges to their interpretation, (2) the high (and variable) stromal content, which particularly confounds gene dosage analyses and transcriptome interpretation, (3) the limited availability of human cell-culture-based resources to overcome this problem and (4) the scarcity

of paired primary–metastasis tissues, particularly treatment-naïve ones, for example for evolutionary studies. Here we characterize large mouse PDAC cell culture resources and combine the results with cross-species comparisons and functional studies to unravel molecular principles underlying PDAC evolution and phenotypic diversification.

Genetic landscapes of mouse PDAC

We initially characterized primary PDAC cell cultures from 38 mice expressing *Kras*^{G12D} conditionally in the pancreas (PK mice)^{9,10} using multiplex fluorescence *in situ* hybridization (M-FISH), whole-exome sequencing (WES) and array comparative genomic hybridization (aCGH). We developed a pipeline for WES data analysis allowing mouse–human comparisons using identical parameter settings. A WES study on microdissected human PDAC (reduced stromal ‘contamination’) served as the reference human data set⁶. Somatic mutation calling identified 318 synonymous and 606 non-synonymous mutations in 38 mouse PDACs (mPDACs) (Extended Data Fig. 1a and

¹Center for Translational Cancer Research (TranslaTUM), Technische Universität München, 81675 Munich, Germany. ²Department of Medicine II, Klinikum rechts der Isar, Technische Universität München, 81675 Munich, Germany. ³German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. ⁴Institute of Pathology, Technische Universität München, 81675 Munich, Germany. ⁵The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ⁶Comparative Experimental Pathology, Technische Universität München, 81675 Munich, Germany. ⁷Anthropology & Human Genomics, Department of Biology II, Ludwig-Maximilians Universität, 82152 Martinsried, Germany. ⁸Helmholtz Zentrum München, Research Unit Radiation Cytogenetics, 85764 Neuherberg, Germany. ⁹Department of Veterinary Medicine, University of Cambridge, Cambridge CB3 0ES, UK. ¹⁰Instituto de Medicina Oncológica y Molecular de Asturias (IMOMA), 33193 Oviedo, Spain. ¹¹Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología (IUOPA), Universidad de Oviedo, 33006 Oviedo, Spain. ¹²Institute of Oncology of Asturias (IUOPA), HUCA, Universidad de Oviedo, 33011 Oviedo, Spain. ¹³Nanomaterials and Nanotechnology Research Center (CINN-CSIC), Universidad de Oviedo, 33940 El Entrego, Spain. ¹⁴Medizinische Klinik und Poliklinik II, Klinikum der LMU München-Grosshadern, 81377 Munich, Germany. ¹⁵Instituto de Biomedicina y Biotecnología de Cantabria (UC-CSIC), 39012 Santander, Spain.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

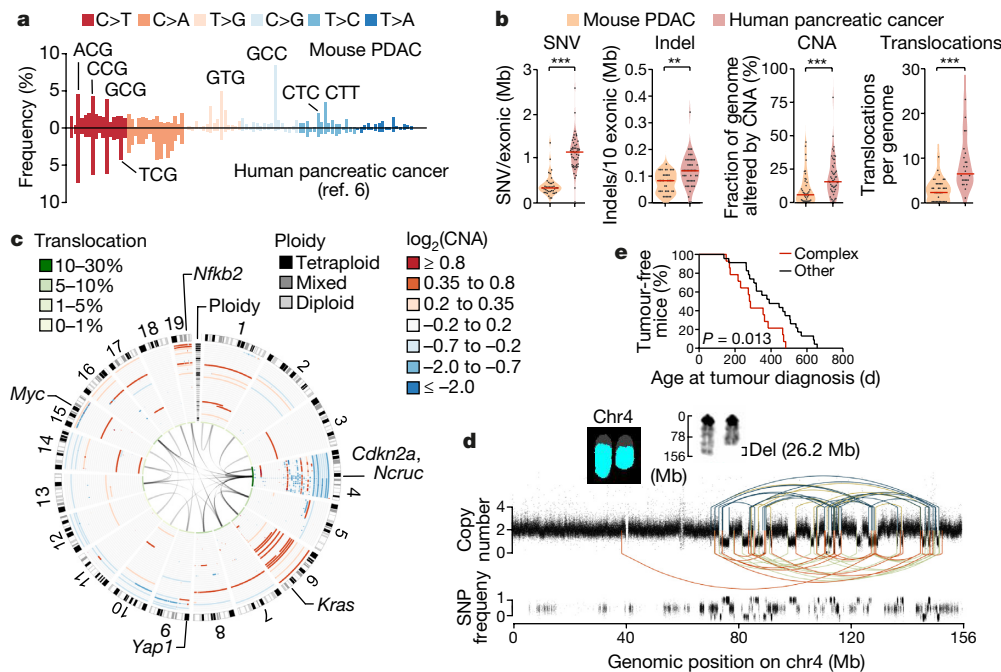


Figure 1 | Genetic landscape of mouse PDAC and comparison with the human disease. **a**, Trinucleotide context-dependent SNV frequencies in mouse ($n = 38$ PK mice) and human PDAC ($n = 51$ patients from ref. 6) derived from WES. **b**, SNV, indel, CNA and translocation burdens by WES, aCGH and M-FISH in PK mice ($n = 38$) and human PDAC ($n = 51$ patients for SNV, indel, CNA (data from ref. 6) and $n = 24$ cell lines for translocations). ** $P = 0.002$, *** $P \leq 0.001$, two-sided Mann–Whitney test; bars, median. **c**, CNAs ploidy and translocations

Supplementary Table 1). Mutational patterns were similar (Fig. 1a, Supplementary Table 2 and Extended Data Fig. 1b–g) but mutational burdens were significantly higher in hPDAC than mPDAC, with 3.3- and 1.5-fold increased median somatic nucleotide variation (SNV) and indel numbers, respectively (Fig. 1b and Supplementary Tables 1 and 3–5). Recurrently altered genes were infrequent in mice. A subset overlapped with recurrently mutated genes in human cancers and/or common insertion sites in pancreas-specific transposon screens^{11–13} (Extended Data Fig. 1a).

Structural changes were also less common in mPDAC (Fig. 1b and Supplementary Tables 6–8). There was, however, substantial variation between cancers, with some mPDACs having only few focal alterations, but others showing extensive changes, including clustered intra-chromosomal alterations, aneuploidy and inter-chromosomal translocations (Fig. 1c and Extended Data Fig. 1h–l). Notably, 34% of tumours (14 out of 38) had complex rearrangements, with ten or more alterations per affected chromosome. The majority (12 out of 14) of such events affected chromosome 4 (chr4), invariably involving *Cdkn2a*. One cancer showed massively rearranged chr15 with high-level *Myc* amplification and another tumour had clustered chr1 rearrangements (Extended Data Fig. 2a–n). These findings reflect the selection of complex rearrangements that affect cancer drivers.

The regularity of oscillating copy-number states in most cancers suggested chromothripsis as the predominant process underlying these complex alterations. Whole-genome sequencing (WGS), followed by rearrangement analysis and computational simulations, confirmed all hallmarks defining the one-off nature of chromothripsis¹⁴, including clustering of breakpoints, regularity of oscillating copy-number states, identical copy-number alteration (CNA) and loss of heterozygosity (LOH) patterns, randomness of DNA segment order/joints and alternating head–tail sequences (Fig. 1d and detailed analyses in Extended Data Fig. 2p–y). In addition, M-FISH confirmed chr4 content loss affecting only one haplotype (Fig. 1d).

in PK mice ($n = 38$), detected by aCGH and M-FISH. Mixed ploidy, $n \geq 3$ diploid/tetraploid cells in ten karyotypes. **d**, Rearrangement graph showing chromosome 4 chromothripsis in mPDAC S821, on the basis of WGS. Haplotype-specific chromosome content loss confirmed by M-FISH ($n = 10$ out of 10 karyotypes). **e**, Age at tumour diagnosis of mice having cancers with ($n = 14$) or without ($n = 23$) complex or clustered chromosomal rearrangements ($n \geq 10$ CNAs per chromosome). Two-sided log-rank test.

Complex rearrangements have been proposed to trigger accelerated evolution of human PDAC¹⁵. The mouse model allows experimental interrogation of this hypothesis because of the ‘synchronized’ nature of tumour initiation (*Kras*^{G12D} mutation). We found that time to tumour development was indeed shorter in animals with *Cdkn2a* loss through catastrophic events (Fig. 1e and Extended Data Fig. 2o). A subset (16%) of complex rearrangements in hPDAC disrupts multiple known tumour suppressors through translocations¹⁵. Chromothripsis-associated chr4 translocations were also frequent in mice (Fig. 1c), although no recurrent translocation partners were found.

Kras^{MUT-iGD} links early progression and metastasis

The most common amplification affected the *Kras* locus (Extended Data Fig. 3a, b), which is also frequently affected in hPDAC^{16,17}. Combined analyses of M-FISH, aCGH and *Kras* mutant/wild-type (WT) allele frequencies revealed four different *Kras*^{G12D} gene dosage ‘states’ (Fig. 2a, Extended Data Fig. 3c–h and Supplementary Table 9): focal gain (*Kras*^{G12D-FG}, 7.9%), arm-level gain (*Kras*^{G12D-AG}, 23.7%), copy-number-neutral loss of wild-type *Kras* (*Kras*^{G12D-LOH}, 36.8%) or no change (*Kras*^{G12D-HET}, 31.6%). Thus, two-thirds of cancers had allelic imbalances causing increased *Kras*^{G12D} gene dosage (hereafter designated *Kras*^{G12D-iGD}), suggesting strong selective pressure for its acquisition. In addition, two *Kras*^{G12D-HET} tumours displayed loss of *Kras*^{WT} mRNA, but high *Kras*^{G12D} expression (blue dots in Fig. 2b), suggesting additional non-genetic mechanisms. Of note, we observed similar *KRAS*^{G12D-iGD} rates and types in human PDAC cell lines (Supplementary Table 10). The increase in gene dosage affects transcriptional output, as *Kras*^{G12D-iGD} mPDAC had higher *Kras*^{G12D} mRNA expression than *Kras*^{G12D-HET} cancers (Fig. 2b and Extended Data Fig. 3i).

Amplification of *Ras/Raf* signalling has been observed at different stages of mammary, intestinal or lung tumorigenesis^{18–21}. To identify the stage of *KRAS*^{MUT-iGD} acquisition in PDAC, we microdissected low-grade human pancreatic intraepithelial neoplasias (hPanIN) from 19 patients and performed amplicon-based deep sequencing of

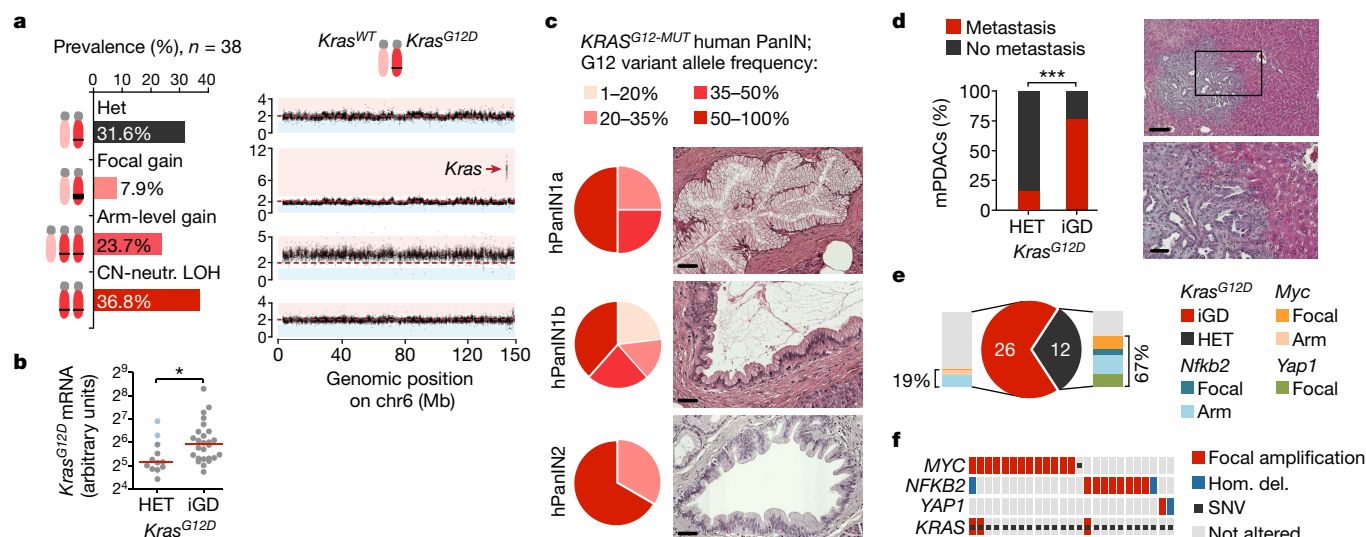


Figure 2 | Mutant *KRAS* gene dosage increase occurs early in PDAC evolution and drives metastasis. **a**, $Kras^{G12D}$ gene dosage 'states' defined by aCGH, WES and M-FISH ($n = 38$ PK mice). Exemplary CNA plot for each state on the right: y axis, copy number. CN-neutr. LOH, copy-number-neutral LOH. **b**, Allele-specific $Kras^{G12D}$ mRNA expression in $Kras^{G12D-iGD}$ ($n = 26$ mice) and $Kras^{G12D-HET}$ mPDACs ($n = 12$ mice) by combined amplicon-based RNA-seq and qRT-PCR. $*P = 0.02$, two-sided Mann-Whitney test; bars, median. **c**, Codon-12 variant allele frequency of microdissected $KRAS^{G12}$ mutant hPanIN ($n = 20$) by amplicon-based deep sequencing. Haematoxylin and eosin stains show histopathological stages of microdissected hPanINs. Scale bars, 50 μ m. **d**, Prevalence of

macro- or micro-metastasis in $Kras^{G12D-HET}$ ($n = 12$) versus $Kras^{G12D-iGD}$ ($n = 26$) mPDACs ($***P = 0.001$, two-sided Fisher's exact test). Liver metastasis, haematoxylin and eosin. Scale bars, 150 μ m (top) and 50 μ m (bottom); rectangle in top panel is expanded in bottom panel. **e**, $Kras^{G12D-HET}$ mPDAC amplify alternative oncogenes (*Myc*, *Nfkb2* or *Yap1*) to intensify partial aspects of Ras downstream signalling. Focal, focal amplification; Arm, arm-level amplification. **f**, Amplification of *MYC*, *NFKB2* or *YAP1* in $KRAS^{G12D-HET}$ human PDAC. Note, these amplified genes can collaborate not only with $KRAS^{G12D-HET}$ but also with $KRAS^{G12D-iGD}$. Hom. del., homozygous deletion. Data from ref. 6.

KRAS exon-2 (Fig. 2c and Supplementary Tables 11 and 12). hPanIN with exon-2-mutated *KRAS* (20 out of 40 hPanINs featured $KRAS^{G12}$ mutations) displayed frequent $KRAS^{G12D-iGD}$. $KRAS^{G12D-iGD}$ allele frequencies greater than 50% in 50%, 38% and 67% of $KRAS$ exon-2-mutated hPanIN1a, hPanIN1b and hPanIN2, respectively. Given that healthy tissue 'contamination' rates in microdissected PanINs ranged between 10% and 60%, $KRAS^{G12D-iGD}$ is likely to be even more frequent. In cases with close to 100% mutant read frequency, *KRAS* interphase FISH excluded false-positive $KRAS^{G12D-iGD}$ arising through chr12 monosomy (Extended Data Fig. 3l–n). Moreover, false-positive $KRAS^{G12D-iGD}$ through cross-'contaminating' hPDAC is excluded because of (1) the large distance of selected hPanINs to associated cancers, (2) distinct *KRAS* mutations in hPanINs and associated cancers or (3) $KRAS^{G12D-iGD}$ in intraductal papillary mucinous neoplasm-related hPanINs without invasive hPDAC. Altogether, these data suggest a critical role of $KRAS^{G12D-iGD}$ in early PDAC progression.

Looking at organ dissemination, we found that $Kras^{G12D-iGD}$ cancers had a markedly increased metastatic potential (odds ratio 16.7; 95% confidence interval 2.8–98.0; Fig. 2d): primary mPDACs with $Kras^{G12D-iGD}$ were mostly metastasized (20 out of 26), whereas $Kras^{G12D-HET}$ mPDACs were predominantly non-metastatic (2 out of 12). Thus, $Kras^{G12D-iGD}$ drives both early progression and metastasis. This dual role explains (1) early PDAC dissemination in humans and mice²² and (2) the high incidence of human PDAC metastasis at diagnosis²³. We also mined published data^{8,24} and invariably found $KRAS^{G12D-iGD}$ in human PDAC metastases. However, because $KRAS^{G12D-iGD}$ is present in the primary tumour (early acquisition) its contribution to metastasis could not be recognized by primary–metastasis comparisons⁸.

Alternative oncogenic gains in $Kras^{G12D-HET}$ tumours

Among the 12 cancers without $Kras^{G12D}$ -dosage gain, two cases had *Myc* amplifications and two had *Yap1* gains (Fig. 2e and Extended Data Fig. 4a–d). *MYC* and *YAP1* are known human oncogenes, amplified in 12% (13 out of 109) and 1% (1 out of 109) of hPDAC, respectively (Fig. 2f). In addition, chr19 gain occurred more frequently in $Kras^{G12D-HET}$

(3 out of 12) than $Kras^{G12D-iGD}$ tumours (4 out of 26), although this difference was not significant. A focal amplification on chr19 contained 20 genes (Extended Data Fig. 4e). Cross-species analyses revealed frequent gains of the syntenic region in hPDAC, with two genes in the minimal peak region: *NFKB2* and *PSD*, both amplified in 7% (8 out of 109) of hPDAC (Extended Data Fig. 4f). *NFKB2* (but not *PSD*) is expressed in human pancreas and hPDAC (Extended Data Fig. 4g, h), suggesting that *Nfkb2* is the target proto-oncogene on mouse chr19. *NFKB2* mediates non-canonical Nfkb signalling. It has not yet been associated with hPDAC, but promotes cell cycle progression *in vitro*²⁵, and knockout of its interaction partner *RelB* impairs PanIN formation in PK mice²⁶. Thus, upon *Kras* mutation, further amplification of partial aspects of *Kras* downstream signalling (*Myc*, *Yap1* or *Nfkb2*) seems sufficient to drive early PDAC progression, whereas strong metastatic potential is linked to amplification of the full *Kras^{G12D}* signalling program.

Evolutionary licensing of oncogenic dosages

The most frequent deletion in mPDAC affected *Cdkn2a* and/or the adjacent non-coding *Cdkn2a*-regulatory region *Ncruc*: 23 *Cdkn2a* ^{Δ HOM}, 4 *Ncruc* ^{Δ HOM}, 10 *Cdkn2a* ^{Δ HET}, 1 *Cdkn2a*^{WT} (chr4 alteration types shown in Fig. 3a, b, Extended Data Fig. 5a–d and Supplementary Table 9). Notably, the majority of *Cdkn2a*/*Ncruc* ^{Δ HOM} cancers were $Kras^{G12D-iGD}$ (23 out of 27) and had high $Kras^{G12D}$ expression. By contrast, *Cdkn2a* ^{Δ HET} or *Cdkn2a*^{WT} cancers were predominantly $Kras^{G12D-HET}$ (8 out of 11) with low $Kras^{G12D}$ expression (Fig. 3c and Extended Data Figs 3j and 5e, f). Accordingly, in microdissected human PDAC data sets⁶, $KRAS^{G12D-iGD}$ variant allele frequencies were higher in *CDKN2A* ^{Δ HOM} than in *CDKN2A* ^{Δ HET/WT} tumours (Fig. 3d and Extended Data Fig. 5g). Thus, *CDKN2A* ^{Δ HOM} deletion and $KRAS^{G12D-iGD}$ are linked, with two possible scenarios: (1) $KRAS^{G12D-iGD}$ occurs first, but induces senescence that prevents progression until *CDKN2A* is lost (as proposed in breast tumorigenesis¹⁸), or (2) $KRAS^{G12D-iGD}$ is tolerated only if preceded by *CDKN2A* deletion.

To resolve the sequence of events, we determined copy-number changes and copy-number-neutral allelic imbalance at *Cdkn2a* and *Kras*

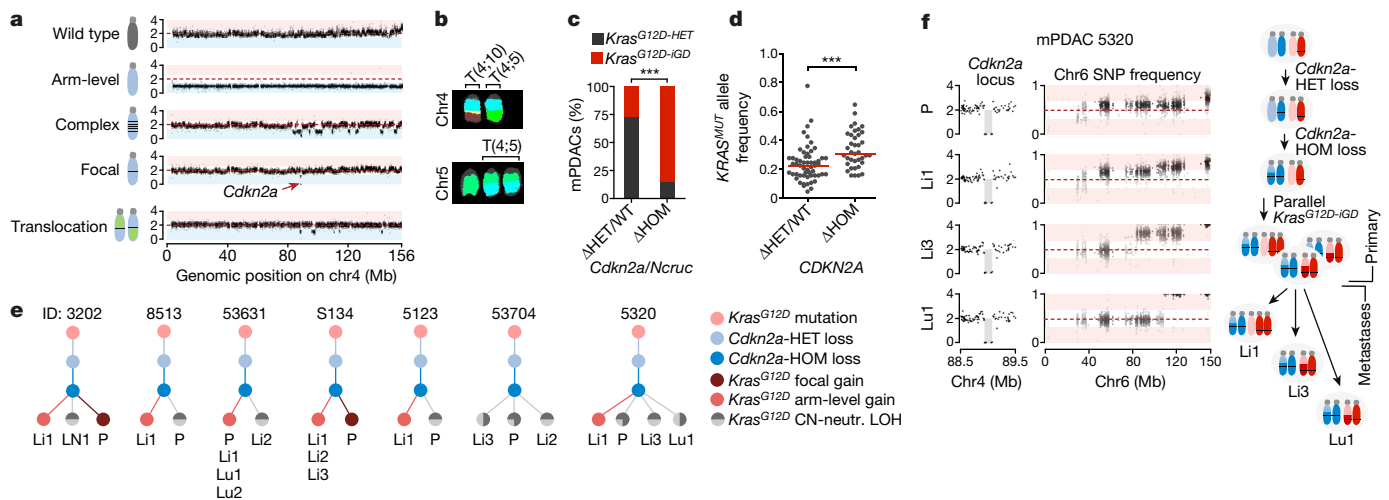


Figure 3 | *Cdkn2a* alteration states dictate distinct evolutionary PDAC trajectories. **a**, Chr4 alteration types involving *Cdkn2a* by aCGH and M-FISH ($n = 38$ PK mice). Complex rearrangements, $n \geq 10$ CNAs per chromosome. Exemplary CNA plots on the right; y axis, copy number. **b**, Translocations affecting chr4/*Cdkn2a* in mPDAC-R1035 by M-FISH (ten out of ten karyotypes). **c**, Prevalence of *Kras*^{G12D-iGD} in mPDAC with homozygously (Δ HOM, $n = 27$) versus wild-type or heterozygously (Δ HET/WT, $n = 11$) deleted *Cdkn2a*/*Ncruc*. *** $P = 0.001$, two-sided Fisher's exact test, odds ratio 15.3, 95% confidence interval 2.8–83.9. **d**, *KRAS* variant allele frequencies in human PDAC with wild-type or heterozygously ($n = 56$) versus homozygously deleted ($n = 38$) *CDKN2A*. Data from ref. 6. *** $P \leq 0.001$, two-sided Mann–Whitney test; bars, median. **e**, Sequential order of *Cdkn2a* and *Kras*^{G12D} alterations. Chr4 and

chr6 CNA and LOH patterns (based on aCGH, WES) of primary mPDACs ($n = 13$ PK mice) and associated metastases ($n = 25$). For 7 mPDACs and 16 associated metastases the order of genetic events (dots) could be reconstructed. Bifurcations, divergent evolution of clones; lines, lengths do not represent evolutionary distances; P, primary tumour; Li/Lu/LN, liver/lung/lymph node metastasis; ID, identifier. **f**, Detailed chr4 and chr6 CNA and LOH profiles for mPDAC 5320 primary tumours and metastases. *Cdkn2a* deletions are identical in all lesions (y axis, copy number). SNP frequency analysis by WES shows distinct chr6 SNP patterns in metastases and a composite picture in the primary, showing convergent evolution of different *Kras*^{G12D-iGD}-gains upon *Cdkn2a* ^{Δ HOM}. Scheme, combined interpretation of WES and aCGH data.

in *Cdkn2a* ^{Δ HOM};*Kras*^{G12D-iGD} mPDACs ($n = 13$) and associated metastases ($n = 25$). We found identical *Cdkn2a* deletions in all 13 primary–metastasis pairs, but discordant chr6 CNA/LOH phenotypes in 7 out of 13 pairs (Fig. 3e). In 6 out of 13 pairs the sequence of *Cdkn2a* loss and *Kras*^{G12D-iGD} could not be reconstructed, either because the somatic nucleotide polymorphism (SNP) density was too low (four cases) or because chr6 profiles in primary–metastasis pairs were identical (two cases). Thus, in all cases with reconstructable sequence, *Cdkn2a* deletion preceded *Kras*^{G12D-iGD} acquisition. For example, mPDAC-53704 (Extended Data Fig. 6) had two liver metastases with identical *Cdkn2a* deletions, but distinct chr6 SNP patterns: one with *Kras*^{G12D-LOH} at distal chr6 (through mitotic recombination) and another affecting the whole chromosome (probably through missegregation). This confirms clonal chr6 diversification and convergent evolution following *Cdkn2a* loss, and explains the primary tumour's gradual chr6 SNP pattern (Extended Data Fig. 6). Figure 3f shows another example: mPDAC 5320 and its three metastases had identical *Cdkn2a* deletions, but distinct chr6 patterns; while liver metastasis-1

had *Kras*^{G12D-AG} (combined interpretation of aCGH and SNP data), liver metastasis-3 and the lung metastasis had distinct *Kras*^{G12D-LOH} events, again showing convergent evolution of *Kras* allelic imbalance and explaining the primary tumour's composite SNP pattern (Fig. 3f).

These results reveal several evolutionary principles in PDAC. First, *Kras*^{G12D-iGD} is contingent on *Cdkn2a* ^{Δ HOM} inactivation. Second, *Myc*, *Yap1* or *Nfkb2* amplifications can occur in a *Cdkn2a* ^{Δ HET} context, suggesting context-dependent *Cdkn2a* haploinsufficiency. Of note, only one cancer was *Cdkn2a*^{WT}. Third, evolution of multiple independent *Kras*^{G12D} gains in *Cdkn2a* ^{Δ HOM} cancers demonstrates functional convergence towards *Kras*^{G12D-iGD} acquisition upon homozygous *Cdkn2a* loss.

To provide *in vivo* evidence for functional convergence in *Cdkn2a* ^{Δ HOM} contexts, we generated mice with pancreas-specific *Kras*^{G12D}-expression and *Cdkn2a*-deletion (PKC). We found *Kras*^{G12D} in 100% (16 out of 16) of PKC tumours (Fig. 4 and Supplementary Table 9), confirming that *Kras*^{G12D-iGD} acquisition is the preferred evolutionary route upon homozygous *Cdkn2a* loss.

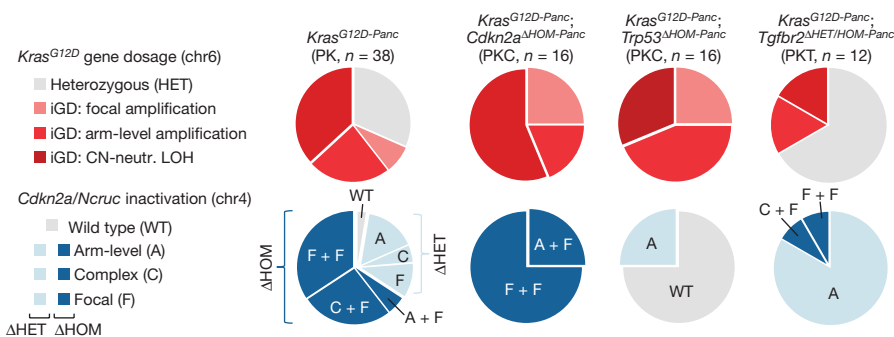


Figure 4 | Defined allelic states and/or combinations of hallmark PDAC tumour suppressor alterations license oncogenic dosage variation. Types and frequencies of *Kras*^{G12D} gene dosage and *Cdkn2a* inactivation, defined by aCGH and amplicon-based *Kras*^{G12D} sequencing in PDAC

mouse models expressing pancreas-specific *Kras*^{G12D} alone (PK) or in combination with engineered *Cdkn2a* ^{Δ HOM} (PKC), *Trp53* ^{Δ HOM} (PKP) or *Tgfb2* ^{Δ HET/HOM} (PKT) inactivation.

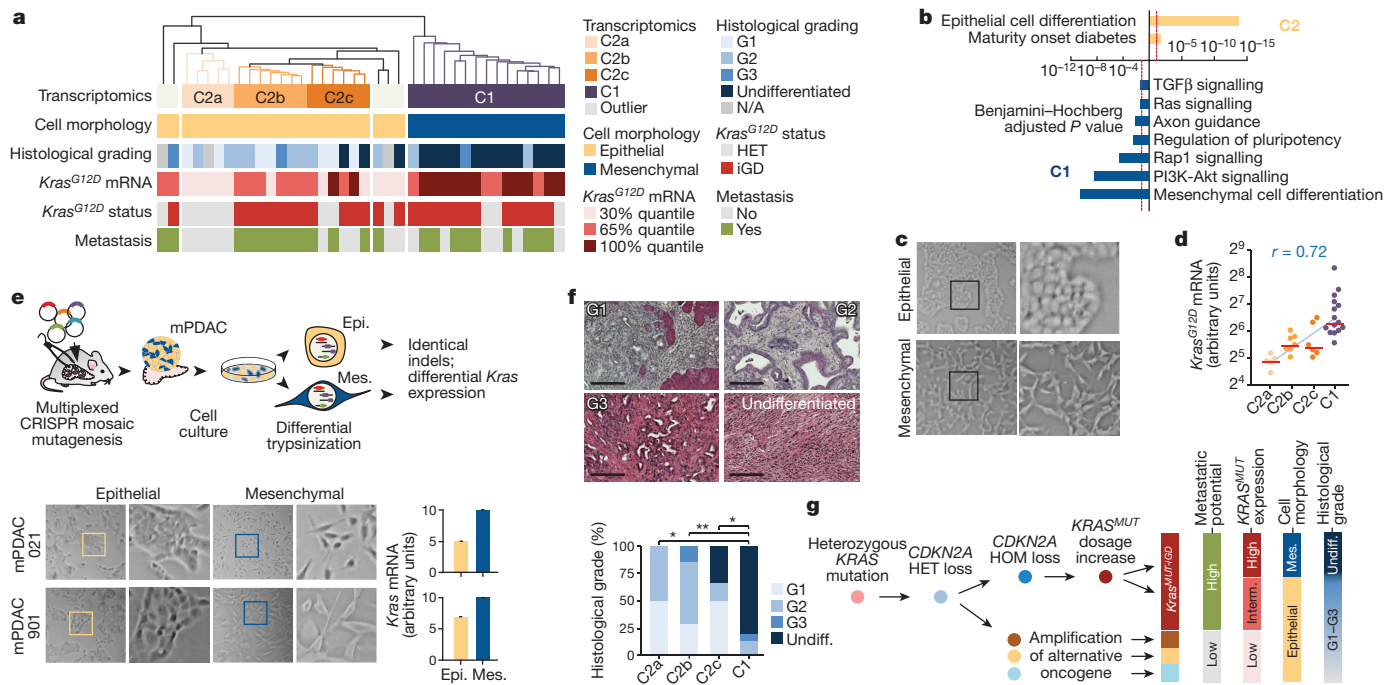


Figure 5 | Integrative analyses of PDAC genomics, transcriptomics, cellular phenotypes and histopathologies link molecular, morphological and clinical disease characteristics. **a**, Unbiased hierarchical clustering of primary mPDAC culture transcriptomes (PK mice). Cell morphology, histopathological grading, *Kras*^{G12D} mRNA expression, genetic *Kras*^{G12D} status and presence or absence of metastasis integrated below. **b**, Selected gene sets from gene set enrichment analysis of clusters C2 versus C1 (full list in Supplementary Tables 13 and 14). **c**, mPDAC cultures with mesenchymal or epithelial morphology from clusters C1 and C2, respectively. Magnification $\times 100$; squares in left panels are expanded on right. **d**, *Kras*^{G12D}-allele-specific mRNA levels in mPDAC transcriptional clusters, combined amplicon-based RNA-seq and qRT-PCR (C2a/b/c/C1, $n = 5/7/6/15$ mice). $P = 1.9 \times 10^{-6}$, two-sided Pearson correlation; bars, median. **e**, CRISPR-Cas9-mediated multiplexed somatic inactivation of PDAC-relevant tumour suppressors by electroporation-based transfection to achieve low-frequency mosaicism and clonal tumour outgrowth. Differential trypsinization separates epithelial (Epi.) and mesenchymal (Mes.) cells in mPDACs with mixed morphologies (magnification $\times 100$; squares in left panels are expanded on right). CRISPR-Cas9-induced indel signatures are identical in epithelial-mesenchymal pairs (Extended Data Fig. 8), indicating common cells of

Another hPDAC hallmark is *TP53* inactivation²⁷. The analysis of cancers from *Kras*^{G12D-Panc}; *Trp53* ^{Δ Panc} (PKP) mice revealed ubiquitous *Kras*^{G12D-iGD} (16 out of 16) (Fig. 4 and Supplementary Table 9). Thus, *Trp53*-loss (like *Cdkn2a* ^{Δ HOM} alteration) predisposes tumours to *Kras*^{G12D-iGD} acquisition (also reflected in hPDAC; Extended Data Fig. 5g). Comparisons of PK, PKC and PKP tumours revealed higher CNA numbers and a tendency to amplify *Kras*^{G12D} through arm-level gain (trisomy) in PKP, whereas copy-number-neutral LOH predominates in PKC (Fig. 4 and Extended Data Fig. 5h). Of note, PKP and PKC tumours did not have complex chr4 rearrangements, confirming that chromothripsis in PK cancers results from natural selection for *Cdkn2a* inactivation (Fig. 4).

To address the role of the TGF β -pathway, we characterized *Kras*^{G12D-Panc}; *Tgfr2* ^{Δ Panc} (PKT) mice (Fig. 4 and Supplementary Table 9). Strikingly, all PKT tumours ($n = 12$) had *Cdkn2a* alterations: two cancers were *Cdkn2a* ^{Δ HOM}/*Kras*^{G12D-iGD}, and ten were *Cdkn2a* ^{Δ HET} and predominantly *Kras*^{G12D-HET} (eight out of ten). Overall, the prevalence of *Kras*^{G12D-iGD} was significantly lower in PKT (4 out of 12) than PK mice (26 out of 38) ($P = 0.04$, Fisher's exact test, odds ratio 0.23, 95% confidence interval 0.06–0.92). *Kras*^{G12D-HET} cancers had frequent alternative oncogenic gains (*Nfkb2*/chr19 trisomy in four out of eight PKT mice), similar to

origin. Total *Kras* mRNA levels in epithelial-mesenchymal pairs (qRT-PCR, normalized to *Gapdh*, $n = 2$ technical replicates). Bars, mean; error bars, s.e.m. **f**, mPDAC histopathological grading in transcriptional clusters (C2a/b/c/C1, $n = 4/7/6/15$, single section per mPDAC). Representative sections (haematoxylin and eosin) shown. *Benjamini-Hochberg adjusted $P \leq 0.05$, ** $P = 0.005$; two-sided Fisher's exact test; scale bars, 150 μ m. **g**, Simplified model of PDAC evolution reconciling molecular, morphological and clinical disease characteristics. *Kras*^{G12D-iGD} gain or alternative oncogenic amplifications (*Myc*, *Yap1* or *Nfkb2*) are critical for early disease progression. Different oncogenic gains and dosages evolve along distinct evolutionary routes, licensed by defined allelic states (heterozygous or homozygous) and/or combinations of hallmark tumour suppressor alterations. For simplicity, only the prototype tumour suppressor gene *CDKN2A* is shown. Not visualized: *TP53* ^{Δ HOM} loss, also promoting *KRAS*^{MUT-iGD}, or *TGFBR2* ^{Δ HET/HOM} inactivation, supporting evolution through *CDKN2A*^{HET}/*KRAS*^{MUT-HET} alternative trajectories. Depicted trajectories are typical, but not completely exclusive; for example, *MYC* or *NFKB2* amplifications, which drive *KRAS*^{MUT-HET} cancers, can also cooperate with *KRAS*^{MUT-iGD}. Major aspects of a cancer's biology and phenotype are linked to differential evolution.

Kras^{G12D-HET} cancers in the PK cohort. Thus, unlike *Trp53* ^{Δ HOM} or *Cdkn2a* ^{Δ HOM} alterations, which license *Kras*^{G12D-iGD}-acquisition, *Tgfr2* alterations facilitate the alternative route with *Cdkn2a* haploinsufficiency.

Altogether, these data show that evolutionary contingencies and convergence shape early tumorigenesis: different tumour suppressor genes or pathways (*Cdkn2a*, *Trp53*, *Tgf β*), their alteration types (Δ HOM or Δ HET) or their combinations (for example, *Cdkn2a* ^{Δ HET} and *Tgfr2* ^{Δ HET}) direct evolution into different trajectories by licensing distinct types and extents of oncogenic dosage gains.

Integrating genomes, transcriptomes, phenotypes

Unbiased hierarchical clustering of RNA sequencing (RNA-seq) data from mPDAC cell cultures (PK cohort) revealed two clusters, C1 and C2, with three sub-clusters within C2 (Fig. 5a). Pathway analyses identified 'epithelial cell differentiation' as the top C2 Gene Ontology term, whereas 'mesenchymal cell differentiation' was defining C1 (Fig. 5a, b and Supplementary Tables 13 and 14). Notably, all C1 cell lines showed mesenchymal cell morphology, while C2 lines were invariably epithelial (Fig. 5a, c).

Previous studies classified human pancreatic cancer on the basis of transcriptional profiles^{7,28,29}. Unbiased hierarchical clustering with

published classifiers shows large overlaps of subtypes proposed by Bailey *et al.*⁷ and Moffitt *et al.*²⁹ with the initially proposed three subtypes of Collisson *et al.*²⁸: classical, exocrine-like, quasimesenchymal. One exception is the lacking exocrine-like signature in the classification of Moffitt *et al.*²⁹, which was proposed to be an artefact of acinar cell ‘contamination’ (details in Extended Data Fig. 7a–d). The Collisson classifier²⁸ separates human PDAC cell lines into two subtypes (classical and quasimesenchymal; Extended Data Fig. 7e) and mouse PDAC cell lines into three subtypes: classical-equivalent, quasimesenchymal-equivalent (both in epithelial C2) and the mesenchymal M subtype (C1) (Extended Data Fig. 7f). The equivalent of the mouse mesenchymal M subtype with the strong epithelial–mesenchymal transition (EMT) signature has not been described in human cell lines so far, reflecting underrepresentation of mesenchymal phenotypes in human cell line collections (see also Extended Data Fig. 7g, h). As described below, however, mesenchymal mPDACs in C1 represent human pancreatic carcinomas with a pronounced EMT signature and undifferentiated histology.

C1 shows strong gene set enrichment for Ras downstream signalling pathways (Fig. 5b and Supplementary Tables 13 and 14). This cannot be explained by the genetic *Kras* status alone: only C2a is *Kras*^{G12D-HET}, whereas C2b, C2c and C1 are mostly *Kras*^{G12D-IGD}. However, integration of *Kras*^{G12D} expression revealed its gradual increase from C2a to C2b/c and further substantial elevation in C1 (Fig. 5d and Extended Data Fig. 3k). Thus, the mesenchymal phenotype is associated with *Kras*^{G12D} expression above a certain threshold.

To study this association further, we induced clonal PDACs by CRISPR–Cas9 somatic mutagenesis³⁰ in PK mice (Fig. 5e), screened for the simultaneous presence of epithelial and mesenchymal cells, and separated and enriched either phenotype by differential trypsinization. Two such cancers were identified. In each case, indel patterns of epithelial–mesenchymal pairs were identical (Extended Data Fig. 8a, b), showing (1) common clonal origin of epithelial and mesenchymal cells and (2) independence of epithelial and mesenchymal phenotypes from CRISPR–Cas9-induced TSG alterations. Notably, however, mPDAC 021 had *Kras*^{G12D-IGD}, elevated *Kras* expression and downstream pathway activation in mesenchymal, but not epithelial, cells. In mPDAC 901, both clones were *Kras*^{G12D-HET}, but mesenchymal cells had increased *Kras* expression, supporting a role of *Kras*^{G12D} dosage variation in shaping cellular phenotypes (Fig. 5e, Extended Data Fig. 8c, d and Supplementary Table 15). Moreover, *KRAS*^{G12D} overexpression in hPDAC cell lines induced an EMT signature, with vimentin upregulation and E-cadherin repression (Extended Data Fig. 8e–g and Supplementary Table 16).

PDAC histology revealed a striking association with transcriptome clusters (Fig. 5a, f). Histopathological grade scores increased from C2a to C2b/c and C1, with C2a being well- or moderately differentiated (G1, G2) and C1 being almost exclusively undifferentiated. Undifferentiated cancers are typically advanced and therefore underrepresented (1–3%) in human surgical series or cell line collections, but autopsy series reported up to 16% hPDACs with at least focal undifferentiated components^{31,32}. Dedifferentiation can occur during disease progression or be triggered by treatment. It is associated with poor prognosis^{32,33}, which is also reflected in mice (Extended Data Fig. 9a). Our results link this aggressive PDAC subtype with the highest *Kras*^{G12D} expression levels and Ras-related transcriptional programs (Fig. 5b, d and Supplementary Table 13). We also screened human transcriptome data (Australian pancreatic cancer cohort of the International Cancer Genome Consortium (ICGC PACA-AU)) for undifferentiated pancreatic carcinomas and performed unbiased hierarchical clustering of differentially regulated genes in undifferentiated cancers (Extended Data Fig. 9b). Of note, undifferentiated human pancreatic carcinomas are characterized by reduced expression of genes involved in ‘epithelial’ (cluster-2) or ‘squamous differentiation’ (cluster-1), and a strong upregulation of genes in cluster-3, containing gene sets enriched for EMT and Ras

downstream signalling (Extended Data Fig. 9b–d and Supplementary Tables 17 and 18).

We exploited the mouse to address complex questions, including cell-based resources (overcoming the stroma richness of human PDAC), primary–metastasis resources (phylogenetic tracking, evolution) and *in vivo* modelling (proof-of-concept functional studies). In addition, discoveries were facilitated by the relatively low complexity of mouse PDAC genomes (easier interpretation). Notably, a transposon-induced PDAC model¹³ showed that our findings are equally valid in contexts of excessive mutational loads (Extended Data Fig. 10 and Supplementary Table 19).

Conclusions

Our study proposes a comprehensive conceptual framework for molecular PDAC evolution and phenotypic diversification. It describes evolutionary trajectories, identifies their genetic hallmarks and shows how oncogenic dosage variation is differentially licensed along individual routes by the three major PDAC tumour suppressive pathways to control critical disease characteristics, including early progression, histopathology, metastasis, cellular plasticity and clinical behaviour (Fig. 5g). *RAS* gene mutations affect more than 30% of human cancers, often involving their allelic imbalance. We therefore presume that the principles identified here are relevant far beyond PDAC.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 2 May; accepted 20 December 2017.

Published online 24 January 2018.

- Rahib, L. *et al.* Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res.* **74**, 2913–2921 (2014).
- Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
- Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109–1113 (2010).
- Biankin, A. V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
- Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
- Witkiewicz, A. K. *et al.* Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744 (2015).
- Bailey, P. *et al.* Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
- Makohon-Moore, A. P. *et al.* Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**, 358–366 (2017).
- Jackson, E. L. *et al.* Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev.* **15**, 3243–3248 (2001).
- Schönhuber, N. *et al.* A next-generation dual-recombinase system for time- and host-specific targeting of pancreatic cancer. *Nat. Med.* **20**, 1340–1347 (2014).
- Pérez-Mancera, P. A. *et al.* The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. *Nature* **486**, 266–270 (2012).
- Mann, K. M. *et al.* Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proc. Natl Acad. Sci. USA* **109**, 5934–5941 (2012).
- Rad, R. *et al.* A conditional piggyBac transposition system for genetic screening in mice identifies oncogenic networks in pancreatic cancer. *Nat. Genet.* **47**, 47–56 (2015).
- Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. *Cell* **152**, 1226–1236 (2013).
- Notta, F. *et al.* A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
- Yamada, H. *et al.* Amplifications of both c-Ki-ras with a point mutation and c-myc in a primary pancreatic cancer and its metastatic tumors in lymph nodes. *Jpn. J. Cancer Res.* **77**, 370–375 (1986).
- Heidenblad, M. *et al.* Detailed genomic mapping and expression analyses of 12p amplifications in pancreatic carcinomas reveal a 3.5-Mb target region for amplification. *Genes Chromosom. Cancer* **34**, 211–223 (2002).
- Sarkisian, C. J. *et al.* Dose-dependent oncogene-induced senescence *in vivo* and its evasion during mammary tumorigenesis. *Nat. Cell Biol.* **9**, 493–505 (2007).
- Junttila, M. R. *et al.* Selective activation of p53-mediated tumour suppression in high-grade tumours. *Nature* **468**, 567–571 (2010).

20. Feldser, D. M. *et al.* Stage-specific sensitivity to p53 restoration during lung cancer progression. *Nature* **468**, 572–575 (2010).
21. Rad, R. *et al.* A genetic progression model of Braf(V600E)-induced intestinal tumorigenesis reveals targets for therapeutic intervention. *Cancer Cell* **24**, 15–29 (2013).
22. Rhim, A. D. *et al.* EMT and dissemination precede pancreatic tumor formation. *Cell* **148**, 349–361 (2012).
23. Stathis, A. & Moore, M. J. Advanced pancreatic carcinoma: current treatment and future challenges. *Nat. Rev. Clin. Oncol.* **7**, 163–172 (2010).
24. Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
25. Schneider, G. *et al.* IKK α controls p52/RelB at the skp2 gene promoter to regulate G1- to S-phase progression. *EMBO J.* **25**, 3801–3812 (2006).
26. Hamidi, T. *et al.* Nuclear protein 1 promotes pancreatic cancer development and protects cells from stress by inhibiting apoptosis. *J. Clin. Invest.* **122**, 2092–2103 (2012).
27. Redston, M. S. *et al.* p53 mutations in pancreatic carcinoma and evidence of common involvement of homocopolymer tracts in DNA microdeletions. *Cancer Res.* **54**, 3025–3033 (1994).
28. Collisson, E. A. *et al.* Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **17**, 500–503 (2011).
29. Moffitt, R. A. *et al.* Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **47**, 1168–1178 (2015).
30. Maresch, R. *et al.* Multiplexed pancreatic genome engineering and cancer induction by transfection-based CRISPR/Cas9 delivery in mice. *Nat. Commun.* **7**, 10770 (2016).
31. Morohoshi, T., Held, G. & Klöppel, G. Exocrine pancreatic tumours and their histological classification. A study based on 167 autopsy and 97 surgical cases. *Histopathology* **7**, 645–661 (1983).
32. Iacobuzio-Donahue, C. A. *et al.* DPC4 gene status of the primary carcinoma correlates with patterns of failure in patients with pancreatic cancer. *J. Clin. Oncol.* **27**, 1806–1813 (2009).
33. Winter, J. M. *et al.* Absence of E-cadherin expression distinguishes noncohesive from cohesive pancreatic cancer. *Clin. Cancer Res.* **14**, 412–418 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank the comparative experimental pathology team for discussions, and A. Selmeier, L. Dajka, O. Seelbach, P. Meyer, T. Schmidt, J. Eichinger and T. Stauber for technical assistance as well as M. Reichert for vector constructs. The work was supported by the German Cancer Consortium Joint Funding Program, the Helmholtz Gemeinschaft (PCCC Consortium), the German Research Foundation (SFB1243; A13/A14) and the European Research Council (ERC CoG number 648521).

Author Contributions S.M., D.S., R.R. designed the study; S.M., T.E., R.M., D.S., R.R. interpreted and visualized data; T.E., S.La., M.Zw., M.B. conducted bioinformatic analyses. S.M., T.E., R.M., S.La., M.Zw., I.V. developed bioinformatic analysis strategies; S.M., R.M., M.Zu., T.K., A.S., B.S., J.G., K.Sc., Z.H., A.A., N.S., C.V., L.R. isolated mPDAC cell cultures; S.M., R.M., J.H., K.U. performed genomics with help from R.Ö.; R.Ö., C.Z. conducted RNA-seq; R.B., S.Lo., B.F., S.K., K.St., F.Y. performed cytogenetics; B.K. performed microdissection; B.K., H.-Y.Y., G.K., W.W., K.St. performed pathological assessment; C.Z., S.P., W.E., K.U., I.V. contributed analytical tools; M.Fri., O.M.D., S.E., F.C.-C., J.R., M.I.S., M.Fra., J.M., G.K., R.M.S., J.C., P.L., G.V., W.W., K.St., W.E., G.S., A.B., D.S., R.R. provided resources and critical input; D.S., R.R. supervised the study; S.M., R.R. wrote the manuscript; T.E., R.M., D.S. edited the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to R.R. (roland.rad@tum.de).

METHODS

Primary mPDAC culture preparation. For two-dimensional primary cell culturing, primary tumours or metastases were cut into small pieces and digested for 1–2 h in 200 U ml⁻¹ collagenase II (Worthington) in DMEM medium (Thermo Fisher Scientific) containing 10% fetal calf serum (FCS, Merck) and 1 × penicillin/streptomycin (Thermo Fisher Scientific). After short-term expansion, primary cells were frozen in 10% dimethyl sulfoxide (Roth) and 50% FCS. For all primary culture experiments, culturing medium (DMEM supplemented with 10% FCS and 1 × penicillin/streptomycin) and cultures with fewer than ten passages were used. Primary cultures were routinely tested for mycoplasma contamination by PCR and authenticated by re-genotyping of cell cultures and corresponding mice.

gDNA and RNA isolation. gDNA from mouse primary cell culture pellets was isolated using the DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturer's instructions. Total RNA was isolated with the RNeasy kit (Qiagen) from 60% to 80% confluent primary cell lines cultured in a 10 cm dish in culturing medium without penicillin/streptomycin and immediately transferred into RLT buffer (Qiagen) containing β-mercaptoethanol.

Histology and micro-metastases screening. For histological characterization of mPDACs, specimens 2 μm thick from formalin-fixed paraffin-embedded material were routinely stained with haematoxylin and eosin and submitted to two veterinary pathologists experienced in comparative pancreatic cancer pathology. Histopathological grading was performed with respect to the most recent consensus report of genetically engineered mouse models³⁴. For histopathological examination of micro-metastases, three liver sections (separated by 200 μm) stained with haematoxylin and eosin were screened for metastatic lesions by a veterinary pathologist.

Animal experiments. Mice were maintained on C57Bl/6;129S6/SvEv mixed background and housed under specific-pathogen-free conditions. Female and male mice were randomly submitted to respective tumour cohorts. For the generation of double- or triple-mutants, pancreas-specific Cre lines^{10,35,36} were intercrossed with *Kras*^{G12D-Panc} (PK mice)^{9,10} only, or in addition with *Cdkn2a*^{ΔHOM-Panc} (PKC)³⁷, *Trp53*^{ΔHOM-Panc} (PKP)^{38,39} or *Tgfb2*^{ΔHET-Panc} and *Tgfb2*^{ΔHOM-Panc} (PKT)⁴⁰ mice. Kaplan–Meier survival curves were generated using Prism (GraphPad software version 5.01). If the animal presented a palpable abdominal mass above 1.5 cm, ascites, signs of sickness or a weight loss of more than 15% of body weight, mice were euthanized in compliance with the European guidelines for the care and use of laboratory animals. For necropsy of tumour-bearing mice, the abdominal cavity was macroscopically checked for pancreatic cancer and for metastases at the main metastatic routes (liver, lung, lymph nodes). Animal studies were approved by the Institutional Animal Care and Use Committees of Technische Universität München (Regierung von Oberbayern, Munich, Germany).

Amplicon-based deep sequencing at the *Kras* locus or of *Kras* mRNA. Fifty nanograms of high-quality genomic DNA or reverse-transcribed mRNA (cDNA) were subjected to amplicon-based deep sequencing. Briefly, the *Kras*^{G12D}-mutated locus was amplified using Q5 High-Fidelity DNA Polymerase (New England Biolabs, 40 cycles) and primers with Nextera adaptor overhangs (Supplementary Table 20). In a second Q5 PCR step (15 cycles), Nextera index primers (Illumina) were added. After each PCR step, solid-phase reversible immobilization clean-up (0.8 ×) was performed using an Agencourt AMPure XP kit (Beckman Coulter GmbH). The pooled library was quantified by SYBR Green qPCR (Thermo Fisher Scientific) and a Kapa Biosystems library quantification kit. In total, 8 pM of denatured library (20% spiked PhiX DNA) was sequenced in 300 bp paired-end mode using a MiSeq system (Illumina). Raw reads were mapped to *Kras* reference sequence (Ensemble release GRCh38p4, Genome Reference Consortium). Variant allele frequencies on chr6 at position 145246771 were calculated.

Microdissection of hPanIN and *KRAS*^{G12} status analysis. Nineteen patients (Supplementary Table 11) with or without a history of pancreatic cancer were included in hPanIN lesion analysis, as approved by the Ethics Committee of the Faculty of Medicine of the Technische Universität München. Patients were classified using World Health Organization recommendations and the TNM staging system. Serially cut specimens (10 μm thick) from formalin-fixed paraffin-embedded material were air-dried overnight. Paraffin was removed through short incubation with xylene. Specimens were briefly stained with haematoxylin and kept wet for the microdissection procedure. Individually diagnosed samples were microdissected under an Axio Imager microscope (Zeiss) using a 20-gauge cannula. Pre- and post-sampling microscopic pictures were taken to (1) document dissection performance and (2) re-identify each specimen on the corresponding haematoxylin and eosin-stained slide. gDNA was extracted as described above using MinElute spin columns (Qiagen) for higher sample concentration. Five microlitres of eluted hPanIN gDNA were submitted to amplicon-based deep sequencing of *KRAS* exon-2 for detection of *KRAS*^{G12} hotspot mutations. In brief, two pairs of custom *KRAS* primers (Supplementary Table 20) were used for nested PCR amplification of the corresponding *KRAS*

region. Illumina Nextera primer pairs were used to add sequencing adapters and indices. PCR steps, library quantification and sequencing were performed as described above. Raw reads were mapped to *KRAS* reference sequence (GRCh38, p10). Variant allele frequencies were calculated for *KRAS*^{G12} hotspot mutations (positions 25398284 and 25398285 on chr12).

WGS. One microgram of high-quality gDNA extracted from primary tumour cell line and corresponding tail was sheared on a Covaris M220 focused ultrasonicator (Covaris) to an approximate fragment size of 500 bp. A library was prepared from 500 ng of fragmented gDNA using a NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) in combination with the adaptor/primer sequences and PCR conditions published previously⁴¹. The final library was quantified by qPCR using a Kapa Biosystems library quantification kit. Equimolar amounts of indexed libraries were denatured and diluted to a final concentration of 1.8 pM and sequenced in 300 bp paired-end mode on a NextSeq 550 system (Illumina) to approximately 20 × coverage. Sequencing raw data were converted to fastq format using bcl2fastq software version 2.18.0.12 (Illumina). Raw reads were trimmed with Trimmomatic version 0.36 to preserve an average base Phred quality of 25, mapped using the BWA-MEM algorithm version 0.7.12 with alternative contig handling and mapped to GRCh38.p5 reference genome.

Inference of chromothripsis. For estimation of copy-number states, Bioconductor HMMcopy package 1.16.0 was used followed by segmentation with Bioconductor DNACopy package 1.48.0. For LOH analysis, variant positions in controls and tumours were computed with samtools mpileup version 1.3.1. Only positions in regions with mapping quality of 60 and an average phredscore of 20 were considered for further analysis. Furthermore, positions harbouring strand bias and variant allele frequencies less than 20% and above 85% in the control were excluded as they were probably homozygous in the germline. The minimal cutoff coverage for a given polymorphic position in the control was set to eight reads. Segmental duplications (University of California, Santa Cruz (UCSC) Genome Browser) and regions with mouse line specific variation (Mouse Genomes Project, REL-1505) were excluded. For this set of SNPs, the difference of frequencies between tumour and control samples were calculated. DELLY version 0.7.6 was used for calling structural variations. Structural variation classes were defined according to DELLY callings: deletion-type (3'-to-5'), duplication-type (5'-to-3') and inversion-type (5'-to-5' and 3'-to-3'). The predicted rearrangements were merged and filtered on the basis of variant frequency, mapping quality and the distance between two connected breakpoints. The existence of chromothripsis was tested by applying the six hallmark criteria proposed in ref. 14. Clustering of structural variation breakpoints was tested using a χ² goodness-of-fit test. The regularity of oscillating copy-number states in the chromothriptic model was compared with a virtual chromosome generated by a Monte Carlo simulation, as described in ref. 42. For each distinct number of breakpoints, 100 simulation runs were completed and mean values as well as 95% confidence intervals were calculated. Interspersed loss and retention of heterozygosity was analysed by calculating the Jaccard index between heterozygously deleted segments and regions comprising LOH and SNP information. The randomness of observed DNA segment order was tested using a Monte Carlo simulation as described in ref. 14. The uniform distribution of structural variation types was tested using a χ² goodness-of-fit test. The Wald–Wolfowitz runs test as implemented in R package randtests 1.0 was performed for testing right-sided against the null hypothesis of randomly distributed 5'-to-3' breakpoint joints sequences.

FISH analyses. For the analysis of copy-number status or large structural alterations of human and mouse primary cell lines, M-FISH was performed as described before⁴³. For *KRAS* gene detection in hPanIN specimens, a ZytoLight SPEC *KRAS*/CEN12 Dual Colour Probe kit was used according to the manufacturer's instructions (ZytoVision). In brief, 2 μm formalin-fixed paraffin-embedded specimens were deparaffinized, pre-incubated in CC2 buffer (at 95 °C for 24 min), treated with pepsin solution (at 37 °C for 8 min) and denatured by a heat treatment step at 80 °C for 8 min on an automated Discovery XT system (Ventana Medical Systems). *KRAS*/CEN12 dual colour probe hybridization was performed by co-denaturing at 75 °C for 10 min and by incubating at 37 °C overnight in a ThermoBrite system (Abbott Laboratories). Slides were washed, nuclei stained with 4',6-diamidino-2-phenylindole (DAPI), covered in an antifade mounting medium and by a coverslip, and stored for confocal laser scanning microscopy analyses at 4 °C in the dark. ZytoLight SPEC *KRAS*/CEN12 Dual Colour Probes and DAPI nuclear stain (excitation/emission: DAPI, 405 nm/415–490 nm; ZyGreen, 503 nm/510–540 nm; ZyOrange, 547 nm/560–650 nm) were detected by confocal laser scanning microscopy using a Leica TCS SP8; DMI8 CS microscope equipped with a 63 ×/1.4 oil immersion objective (Leica). Images (z-stacks, covering the whole nucleus) with a magnification factor of 3 and a frame size of 2,048 pixels × 2,048 pixels were collected. Generated images were processed using Huygens Essential software (Scientific Volume Imaging) for deconvolution, then merged and maximum projections were converted with Leica LAS X software.

aCGH analysis. Agilent oligonucleotide aCGH (SurePrint G3 Mouse CGH 240K or custom 60K microarray) was performed according to the manufacturer's instructions. Agilent Genomic Workbench software version 7.0.4.0 was used for aCGH data preprocessing. Legacy centralization option was used for re-centralization of raw log ratios to the most common ploidy state. The ADM-2 algorithm was applied for aberration calling. Segments coordinates were reported for GRCh37 reference genome. Aberrations on chromosome 6 between positions 148719747 and 149503634 were excluded from further downstream analysis as this region probably resulted from an artefact. Normalized and curated data were imported into R.

WES analysis. Coding exons were enriched by whole-exome pull-down using an Agilent SureSelect Mouse Exon Kit according to the manufacturer's instructions and sequenced on an Illumina HiSeq2000 system. Before mapping, raw sequencing reads were trimmed using Trimmomatic version 0.33. Leading and trailing bases with Phred scores below 25 and reads shorter than 50 nucleotides were removed. In addition, the average base quality within a sliding window of 10 nucleotides should be above 25 to keep the read for further downstream analysis. Reads were aligned to the GRCh38.p3 reference genome using BWA-MEM 0.7.12 with default settings. PCR duplicates were marked with Picard tools version 1.130 and realignment around indels was performed with GATK toolkit version 3.4.46. Mutect version 1.1.7 was used for calling somatic mutations with default settings. Potential somatic events were filtered for SNPs by excluding SNVs which were listed in release 1505 of the Mouse Genome Project SNP database⁴⁴. Somatic point mutations were included in the final list, if the read coverage for each position was at least 10 in both control and tumour, variant frequency was at least 10% and read count supporting the variant nucleotide was at least 3 in the tumour sample and equal to 0 in the control. Further, SNVs marked as strand or PCR bias artefacts by 'DKFZBiasFilter' (<https://github.com/eislab/DKFZBiasFilter>, using default settings) or with a FOXOG-Score of 1 were excluded. Annotation of somatic events was conducted with SNPeff version 4.1. SNVs causing variation in splice sites or upstream/downstream of genes were excluded from further analysis. Indels were detected with Pindel⁴⁵. For each potential indel, the read coverage was re-calculated using bedtools version 2.17.0. Criteria for further downstream processing were variant frequency at least 10% in tumour and equal to 0% in control; and total coverage at the altered position in both control and tumour at least 20. LOH analysis was conducted as described in the section above on Inference of chromothripsis.

WES data analysis from hPDAC. Mapped BAM files from ref. 6 were downloaded from the Sequence Read Archive (accession number PRJNA278883), approved by the Ethics Committee of the Faculty of Medicine of the Technische Universität München. Further downstream analysis was performed as described above. SNPs were filtered by excluding variants with an alternative allele frequency of at least 1% in the 1000 Genomes Project, as listed in dbSNP build 146. All available VCF files from the pancreatic adenocarcinoma cohort of The Cancer Genome Atlas generated by Mutect2 were downloaded from the National Institutes of Health Genomic Data Commons data. Downstream processing was performed as described above (PCR and strand bias marking by DKFZBiasFilter was not possible using VCF files). SNPs were filtered by excluding variants with an alternative allele frequency of at least 1% in the 1000 Genomes Project, as listed in dbSNP build 142. MAF files from other human pancreatic cancer cohorts were downloaded and included in our analysis: all samples for which WES data were available, as provided in ref. 7; pancreatic cancer cell lines were from the Cancer Cell Line Encyclopedia⁴⁶ and SNV data from pancreatic cancers as analysed in ref. 47. In these cohorts, SNPs were filtered by excluding variants with an alternative allele frequency of at least 1% in the 1000 Genomes Project, as listed in dbSNP build 146. Remaining SNVs were annotated and filtered with SNPeff as described above.

Analysis of mutational signatures. Mutation spectra for each cohort were compared with a list of 21 signatures previously described in ref. 47; signature 1B was excluded from further analyses because of presumed biological similarity to signature 1A. The contribution of each individual signature to the mutation spectrum of each cohort was analysed by using 'deconstructSigs' version 1.8.0.

qRT-PCR analysis. Reverse transcription was performed with random hexamers using 1 µg of total RNA according to the SuperScript II protocol (Thermo Fisher Scientific). Real-time qPCR was performed either with TaqMan qPCR chemistry (Thermo Fisher Scientific) for mouse using *Kras*-specific primers and probes or with SYBR Green master mix (Thermo Fisher Scientific) using primers for human target genes *VIM*, *CDH1* and *MMP1* (Supplementary Table 20). *Gapdh* or *GAPDH* and *PP1A* were used as housekeeping genes for normalization (Supplementary Table 20). qPCR was conducted on a StepOnePlus system (Applied Biosystems). For analyses of mutant *Kras*^{G12D} mRNA levels in mPDACs, first total (wild-type plus mutant) *Kras* mRNA levels were determined using qRT-PCR. Second, the identical cDNA was used for amplicon-based deep sequencing to detect the proportion of mutant to wild-type *Kras* mRNA. Third, the mutant to wild-type *Kras*

mRNA ratio was multiplied by the total *Kras* mRNA level to calculate the mutant *Kras*^{G12D}-specific mRNA level.

RNA-seq analysis. Bulk 3' transcript end RNA-seq (SCR-seq) libraries were prepared as described previously⁴⁸. In brief, RNA was reversely transcribed using oligo-dT primers decorated with sample barcodes, unique molecular identifiers and adapters (Integrated DNA Technologies). cDNA from all samples was pooled and unincorporated primers digested using ExonucleaseI (New England Biolabs). Next, the cDNA pool was amplified with KAPA HiFi ReadyMix (KAPA Biosystems). To obtain sequencing libraries, 0.8 ng of cDNA was tagged and 3' ends amplified with a Nextera XT Kit (Illumina) using a specific primer for the adaptor on the 3' end. The library was paired-end sequenced on a HiSeq1500 with 16 cycles for read 1 to decode sample barcodes and unique molecular identifiers, and 51 cycles on read 2 into the cDNA fragment. For the preparation of the human pancreatic cancer cell line samples, the flow cell binding sites P5 and P7 were exchanged to allow sequencing of the cDNA in read 1 and barcodes and unique molecular identifiers in read 2. Data were processed using the published Drop-seq pipeline (version 1.0)⁴⁹ to generate sample- and gene-wise unique molecular identifier tables. Reference genome (GRCh38) was used for alignment. Transcript and gene definitions were used according to the ENSEMBL annotation release 75. Further analyses were performed with R version 3.2.2. Initial hierarchical clustering (method: complete linkage; distance measure: Euclidian) of samples was performed for the top 10% of variable genes. Bootstrapping was performed to access cluster stability with the pvclust package version 2.0. The four most prominent clusters were selected and differential expression between these clusters was calculated with DESeq2 (ref. 50). A gene was considered to be differentially regulated if the absolute log₂(fold change) was above 0.8 and the adjusted *P* value was no more than 0.05. Gene set enrichment testing was performed with DAVID 6.8 (ref. 51) or the hypergeometric test as implemented on the 'Molecular Signature Database' (MSigDB) version 6.0 homepage (<http://software.broadinstitute.org/gsea/msigdb/annotate>). For all MSigDB analyses, the top 100 enriched terms with a false discovery rate of $P \leq 10^{-4}$ were included. Published PDAC classifier genes²⁸ and the hallmark EMT gene set (downloaded from MSigDB version 5.2 (ref. 52)) were used for sample clusterings (method: Ward; distance measure: Euclidian).

hPDAC subtyping. Normalized RNA-seq data were derived from ref. 7. Samples that were histologically classified as 'PDA-adenosquamous carcinoma' and 'pancreatic ductal adenocarcinoma' were used for hierarchical clustering (method: Ward; distance measure: Euclidian) with classifier gene lists published elsewhere^{28,29}.

Microarray data analysis. The Affymetrix-based CCLE raw data set was downloaded from (Broad-Novartis Cancer Cell Line Encyclopedia, version 2.17). Haematopoietic or lymphoid neoplasms were excluded since (1) the primary interest of our study was solid tumours and (2) the overall gene expression signature of these samples was shown to be very distinct from all other samples in ref. 46. Normalization of the data was performed with RMA. In general, if genes were represented by two or more probe sets, the probe set with the highest mean expression was used for all further microarray data analyses. Mapping between probe set and genes was conducted with the appropriate Bioconductor packages. Target genes for the TP63ΔN network were downloaded from the 'Pathway Interaction Database' (PID)⁵³ and hierarchically clustered (method: Ward; distance measure: Euclidian). Gene set enrichment analysis was conducted with DAVID or MSigDB version 6.0. All following microarray data sets are Illumina-based and were VST-transformed followed by quantile normalization as implemented in lumi⁵⁴. The microarray data set of hPDAC cell lines (accession number GSE17891) was downloaded from the Gene Omnibus Expression database. PDAC classifier genes and EMT hallmark gene set were used as described above. For the comparison of human wild-type pancreatic tissue and hPDAC cell lines, limma⁵⁵ was used for detection of differential expression between groups. Differentially expressed genes were determined with an alpha level threshold of 5%. The PACA-AU ICGC data set was downloaded from <https://dcc.icgc.org/repositories>. Samples that met the following criteria were selected for further analyses: (1) histosubtypes 'PDA-adenosquamous carcinoma' or 'pancreatic ductal adenocarcinoma' with available subtype information from ref. 7 and (2) ICGC World Health Organization grading 'undifferentiated carcinoma'. Only representative samples, as judged by cluster analysis, from this group were selected for downstream analysis. Analysis of variance (ANOVA) was performed across six defined subgroups of pancreatic cancer: (1) undifferentiated pancreatic carcinoma; (2) adenosquamous pancreatic carcinoma; and (3–6) PDAC sub-stratified in pancreatic progenitor, immunogenic, squamous and aberrantly differentiated endocrine exocrine subtypes. Genes with an adjusted $P \leq 0.05$ were hierarchically clustered (method: Ward; distance measure: Manhattan) and the resulting cluster tree was computationally stratified into five sub-clusters. Genes within sub-clusters were used for gene enrichment analysis as described above. Seventeen PK-PB primary cultures established elsewhere¹³ were submitted to RNA extraction and subsequent gene expression profiling analysis on a MouseWG-6

version 2.0 Expression BeadChip (Illumina). The 5% of genes with the highest variability across all samples were used for hierarchical clustering using the Ward method for aggregation of samples. Limma was used as described above. A gene was called differentially expressed if the adjusted *P* value was no more than 0.05 and the log₂(fold change) was at least 0.8.

Quantitative transposon insertion-site sequencing. Transposon integration sites in PK–PB pancreatic cancer cell cultures¹³ were identified using quantitative insertion-site sequencing (QiSeq) followed by bioinformatics analyses, as described earlier⁵⁶. Transposon integration sites supported by at least 20 reads and residing in intragenic regions were counted for the computation of the mutational burden. For the assessment of the *Cdkn2a/Ncruc* inactivation status caused by transposon mutagenesis, only the top hit of each tumour was considered.

Kras^{G12D} induction after lentiviral transduction of hPDAC cell lines. The pINDUCER20 (ref. 57) vector system comprising a puromycin resistance gene was used for doxycycline-inducible *KRAS*^{G12D} overexpression. In brief, cDNAs of oncogenic *KRAS*^{G12D} (CCDS 8702.1, 35G>A) and *GFP* were cloned into the pINDUCER20 lentiviral vector. Stbl3 bacteria (Thermo Fisher Scientific) were chemically transformed and the pDNA sequence was verified. For lentivirus production, HEK293FT cells were transfected using TransIT-LT1 (Mirus Bio LLC) with standard virus packaging plasmids and respective pINDUCER20 vectors by following the manufacturer's recommendations. Virus-containing supernatant was pooled 48 and 72 h after transfection, concentrated by polyethylene glycol 6000 precipitation⁵⁸ and stored at –80 °C after shock-freezing. One hundred thousand HUPT3 (COSMIC ID: COSS907285) and PANC0327 (COSMIC ID: COSS925346) hPDAC cells were transduced in the presence of 1 µg µl^{–1} polybrene and selected with puromycin antibiotic. Target gene expression was induced for the stated time points by the addition of 100 ng µl^{–1} doxycycline into penicillin/streptomycin-free culturing medium. RNA isolation, qRT–PCR and SCRB-seq were performed as described above. For differential gene expression analysis, raw sequencing data were mapped to the human reference genome (GRCh 38p10). Transcript and gene definitions were used according to the ENSEMBL annotation release 87. Group comparisons (*KRAS*^{G12D} versus *GFP*) were conducted with DESeq2. hPDAC cell lines were routinely tested for mycoplasma contamination by PCR and authenticated by mutation genotyping.

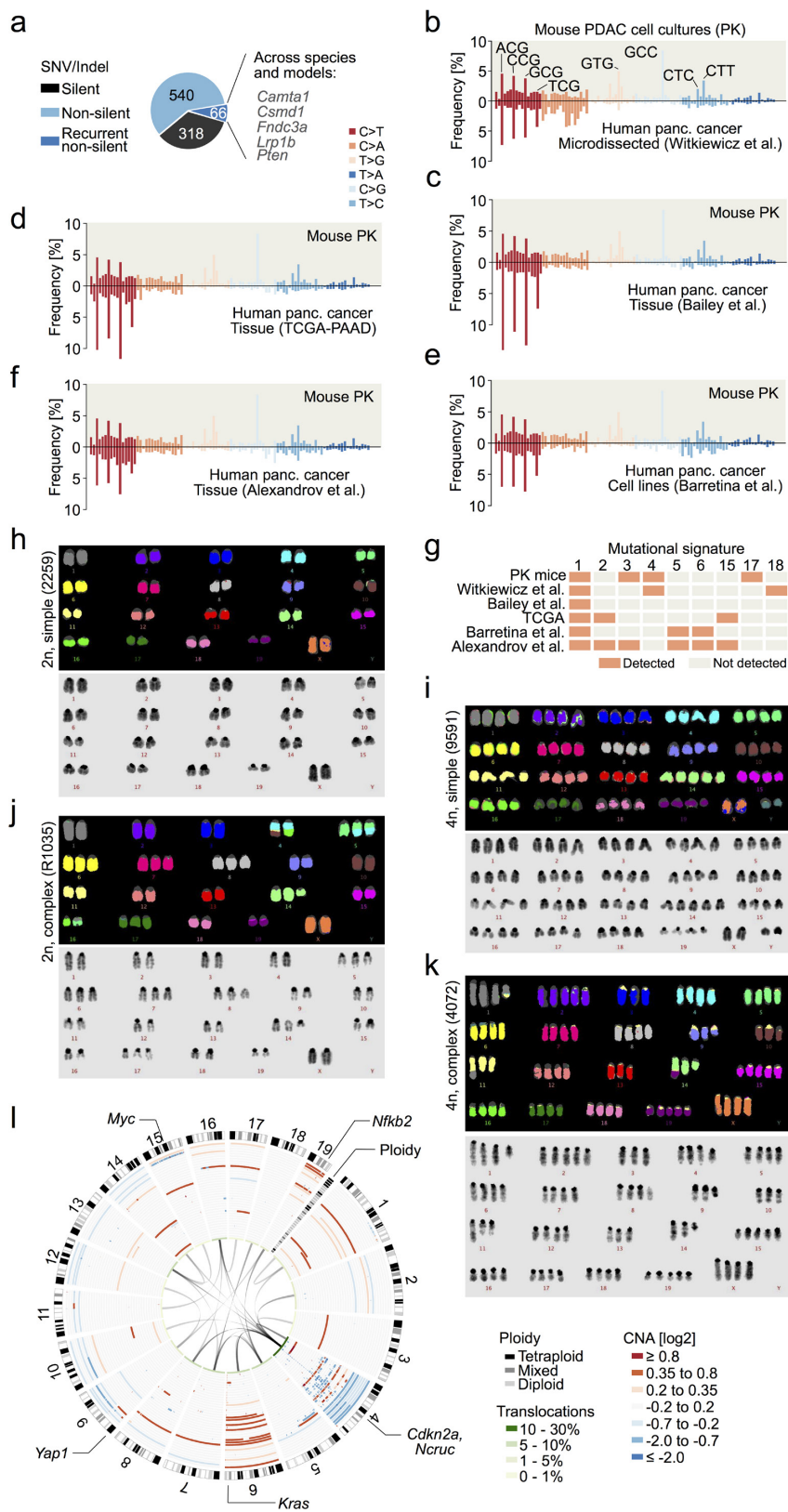
Somatic CRISPR–Cas9 gene editing for tumour clone tracking in mice. Multiplexed gene editing of tumour suppressor genes using CRISPR–Cas9 in the pancreas of PK mice was performed as described elsewhere³⁰. Primary cultures of induced mPDACs were isolated as described above and monitored for the simultaneous presence of epithelial and mesenchymal phenotypes. Enrichment of epithelial and mesenchymal cell morphologies was achieved by differential exposition times to trypsin (Thermo Fisher Scientific). Short-term incubation (2–3 min) at room temperature induced detachment of mesenchymal cells, while epithelial colonies remained adherent. Both cell fractions were subsequently grown to 80% confluency in new flasks. This process was repeated three to six times until homogenous epithelial and mesenchymal cell fractions were enriched. The clonal origin of both phenotypes was confirmed by targeted amplicon-based next-generation sequencing of CRISPR–Cas9-edited loci as described earlier^{30,59}. Analyses of the *Kras* allelic status and mRNA expression were performed as described above.

Statistics and reproducibility. For each experiment, all statistics were performed as indicated in the respective figure and Extended Data figure legends. Statistical testing across all classes was performed to account for multiple testing. Continuous variables were tested for normal distribution. Non-parametric tests were used for non-normally distributed data. Complex statistical techniques are explained in detail elsewhere in the Methods. No animals were excluded from any of the cohorts. The veterinarian pathologists were blinded during histological grading of primary tumours and metastasis screening. The study was of explorative nature. Owing to this study design, previous knowledge of the expected effect size was not available and no power calculations were conducted. The experiments were not randomized.

Code availability. Source code is available from the corresponding author upon reasonable request.

Data availability. Sequence data have been deposited at EBI European Nucleotide Archive under accession number PRJEB23787. Microarray data have been deposited in the Gene Expression Omnibus under accession number GSE107458. All data are available from the corresponding author upon reasonable request.

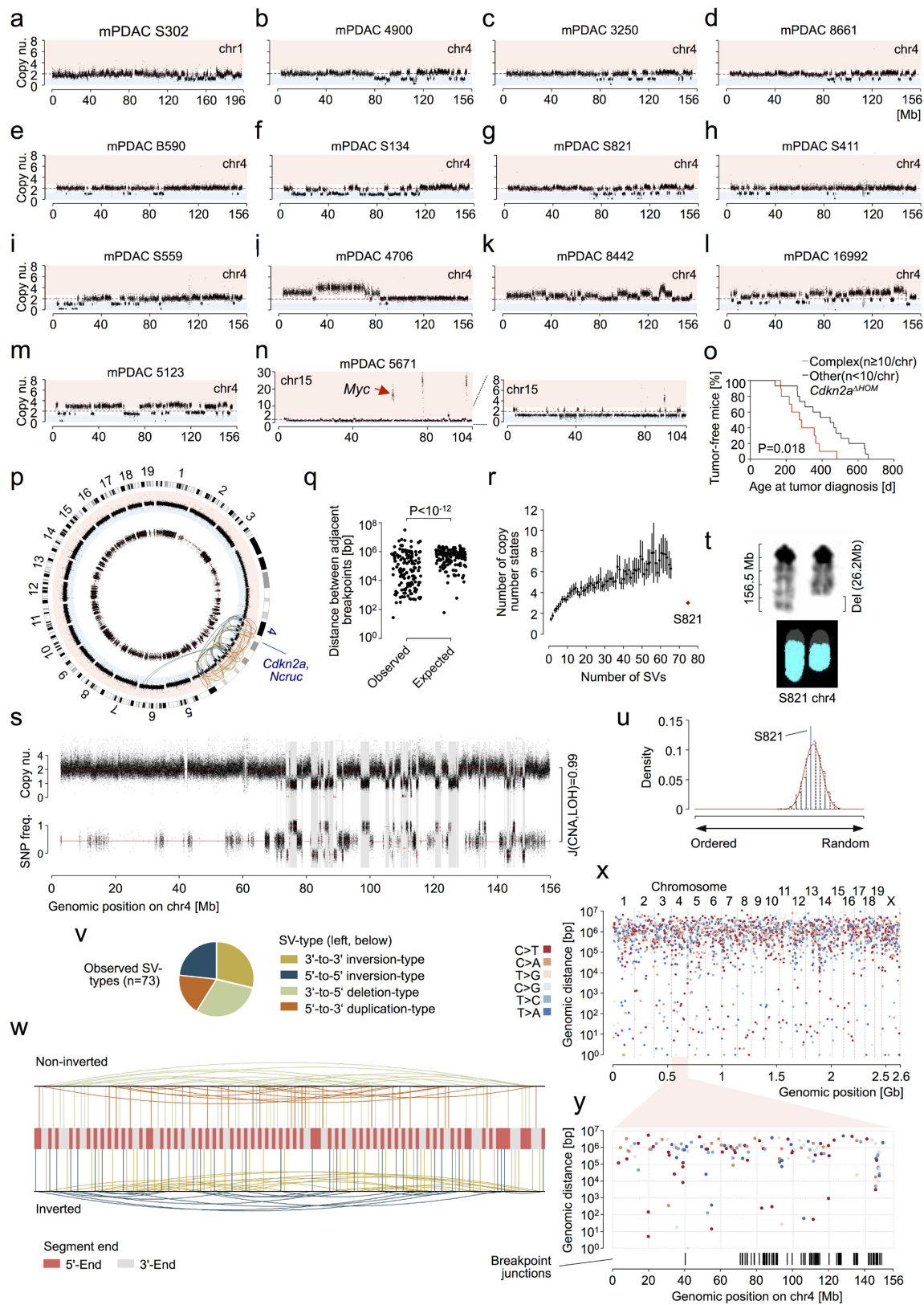
34. Hruban, R. H. *et al.* Pathology of genetically engineered mouse models of pancreatic exocrine cancer: consensus report and recommendations. *Cancer Res.* **66**, 95–106 (2006).
35. Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437–450 (2003).
36. Nakhai, H. *et al.* Ptf1a is essential for the differentiation of GABAergic and glycinergic amacrine cells and horizontal cells in the mouse retina. *Development* **134**, 1151–1160 (2007).
37. Aguirre, A. J. *et al.* Activated Kras and *Ink4a/Arf* deficiency cooperate to produce metastatic pancreatic ductal adenocarcinoma. *Genes Dev.* **17**, 3112–3126 (2003).
38. Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29**, 418–425 (2001).
39. Lee, C. L. *et al.* Generation of primary tumors with Flp recombinase in FRT-flanked p53 mice. *Dis. Model. Mech.* **5**, 397–402 (2012).
40. Chytil, A., Magnuson, M. A., Wright, C. V. & Moses, H. L. Conditional inactivation of the TGF-β type II receptor using Cre:Lox. *Genesis* **32**, 73–75 (2002).
41. Bronner, I. F., Quail, M. A., Turner, D. J. & Swerdlow, H. Improved protocols for Illumina sequencing. *Curr. Protoc. Hum. Genet.* **80**, 18.2.1–18.2.42 (2014).
42. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
43. Jentsch, I., Adler, I. D., Carter, N. P. & Speicher, M. R. Karyotyping mouse chromosomes by multiplex-FISH (M-FISH). *Chromosome Res.* **9**, 211–214 (2001).
44. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
45. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
46. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
47. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
48. Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. & Hellmann, I. The impact of amplification on differential expression analyses by RNA-seq. *Sci. Rep.* **6**, 25533 (2016).
49. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
51. Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44–57 (2009).
52. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
53. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
54. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547–1548 (2008).
55. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
56. Friedrich, M. J. *et al.* Genome-wide transposon screening and quantitative insertion site sequencing for cancer gene discovery in mice. *Nat. Protocols* **12**, 289–309 (2017).
57. Meerbrey, K. L. *et al.* The pINDUCER lentiviral toolkit for inducible RNA interference in vitro and in vivo. *Proc. Natl Acad. Sci. USA* **108**, 3665–3670 (2011).
58. Kutner, R. H., Zhang, X. Y. & Reiser, J. Production, concentration and titration of pseudotyped HIV-1-based lentiviral vectors. *Nat. Protocols* **4**, 495–505 (2009).
59. Weber, J. *et al.* CRISPR/Cas9 somatic multiplex-mutagenesis for high-throughput functional cancer genomics in mice. *Proc. Natl Acad. Sci. USA* **112**, 13982–13987 (2015).
60. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
61. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1 (2013).
62. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
63. Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929–944 (2014).



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Mutational patterns, karyotype complexity and structural alterations in primary PDAC. **a**, Single nucleotide variants (SNVs) and indels in primary PDAC cultures derived from 38 *Kras*^{G12D} (PK) mice, as detected by whole-exome sequencing. Recurrently mutated genes that are frequently altered in human cancers and/or genome-wide pancreas-specific transposon screens are indicated. **b**, Frequency of somatic base substitutions based on trinucleotide context in mouse ($n = 38$ PK mice) and human PDAC ($n = 51$ patients, data used for analysis from ref. 6). **b–f**, Mutation spectra defined by trinucleotide contexts around base substitutions as detected by WES show similar patterns in PK mice ($n = 38$) and in relevant human pancreatic cancer cohorts. Base substitutions were extracted from BAM, VCF or MAF files from (**b**) ref. 6, (**c**) ref. 7, (**d**) the pancreatic adenocarcinoma cohort of The Cancer Genome Atlas, (**e**) ref. 46 and (**f**) ref. 47. Additional information about the analysis of each cohort is provided in Supplementary Table 2. **g**, Mutational signatures in mouse and human pancreatic cancer cohorts. Information on mutational signatures was used from ref. 47, which identified 21 mutational signatures operative in human cancer. The ‘deconstructSigs’ tool was used to determine the composition of the given set of 21 mutational signatures in each pancreatic cancer cohort. Extraction of mutational signatures strongly depends on SNV load per tumour. Owing to the low mutational burden of mPDACs from PK mice (median of 18 SNVs per tumour as detected by WES), the analyses of mutational signatures could not be performed at the level of individual tumours. We therefore investigated the contribution of each of the 21 mutational signatures to the SNV spectrum at the cohort level (see Methods). Signature 1, reflecting age-associated C>T transversions at NCG trinucleotides, was the only signature consistently identifiable in all cohorts of human and mouse pancreatic cancer. Compared with human cohorts, PK mice show C>G substitutions at GCC trinucleotides that cannot be attributed to one of 21 mutational signatures. Note that

mutations at the GCC motif are not a general phenomenon of PDAC from PK mice, since only four samples predominantly contribute to this peak. **h, i**, Representative M-FISH karyotypes with no or few karyotypic changes are shown for a diploid (40 chromosomes) and tetraploid (81 chromosomes) mouse PDAC. Tumour 9591 shows gain of chr14. **j**, Representative karyotype of a complex diploid mPDAC genome with aneuploidy and translocations (46 chromosomes). Both copies of chr4 are involved in translocations: der(4)t(4;10) and der(4)t(4;16), probably affecting *Cdkn2a*. Further structural alterations and copy number changes are +5, der(5)t(4;5)*2, +6, +7, +8, del(9), +14, del(14), der(16)t(5;16), +17. **k**, Representative example of a complex tetraploid mPDAC karyotype (77 chromosomes). Structural alterations are der(1)t(1;11), dic(9;9), der(11)t(1;11) and der(14)t(14;19). Single chromosomal copy number changes are +2, -3, -9, -10, -11, -13, -14, +15 and +19. Del, deletion; der, derivative chromosome; dic, dicentric chromosome; t, translocation; ‘-’, chromosome loss; ‘+’, chromosome gain. **l**, (Extension to Fig. 1c.) Circos plot shows CNAs assessed by aCGH as well as translocations and ploidy states detected by M-FISH in 38 primary PDACs derived from PK mice ($n = 38$). CNAs for each mPDAC are displayed as \log_2 (difference from tail control). Frequencies of translocations per chromosome are indicated in green in the inner circle of the graph. Connecting lines indicate individual translocations and involved chromosomes. On chr4, genomic alterations frequently involve *Cdkn2a* or *Ncruc*, a non-coding regulatory region upstream of *Cdkn2a* (27 out of 38 cancers with homozygous and 10 out of 38 with heterozygous inactivation of *Cdkn2a* and/or *Ncruc*). Only one cancer remained *Cdkn2a*^{WT}. The target of copy number changes on chr6 is *Kras*^{G12D}, either through arm level gain or focal amplification. In addition, primary mPDAC of PK mice exhibited recurrent genetic amplifications affecting other known oncogenes, such as *Myc* or *Yap1*, or *Nfkb2*, a novel oncogenic PDAC driver identified in this study (see also Fig. 2e, f and Extended Data Figure 4).

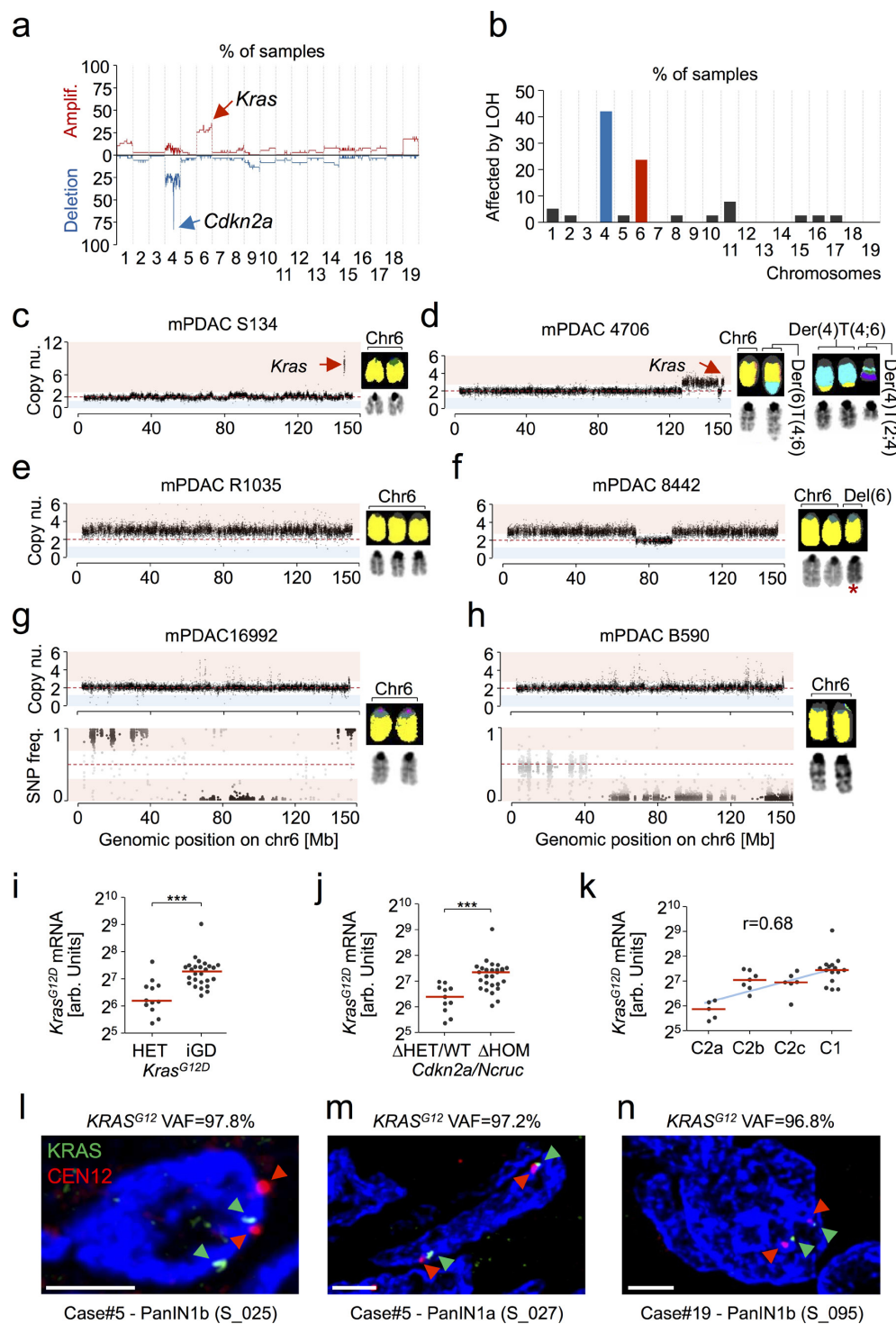


Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | Characterization of complex rearrangements in PDAC from PK mice and statistical inference of chromothripsis on the basis of WGS. a–n, Copy-number profiles of chromosomes with complex rearrangements (defined as $n \geq 10$ CNAs per chromosome) from primary mPDAC cell cultures as detected by aCGH. A total of 14 mPDACs had chromosomes with complex rearrangements. a–i, Nine primary mPDACs show copy-number patterns characterized by heterozygous deletions and oscillation of copy number around few states, indicating chromothripsis as the underlying mechanism. g, mPDAC-S821 was subjected to WGS for the inference of chromothripsis using previously established criteria¹⁴ (see Fig. 1d and Extended Data Fig. 2p–w).

j–m, Four primary mPDACs showed complex rearrangements with multiple copy number states on chr4, probably acquired through progressive or sequential rearrangement cycles. n, Cancer 5671 carries a complex rearrangement on chr15 characterized by oscillating copy number states and three prominent focal amplifications, of which one contained the *Myc* oncogene. *Myc* amplification is most probably the result of double minute chromosome formation during chromothriptic rearrangement of chr15. o, Comparison of age at tumour diagnosis in *Cdkn2a*^{ΔHOM} deleted cancers with ($n = 10$) or without ($n = 15$) complex clustered chromosomal rearrangements ($n \geq 10$ CNAs per chromosome). Complex clustered rearrangements are associated with significantly shortened time to tumour diagnosis, indicating accelerated tumour evolution through genetic crisis. Two-sided log-rank test. p, Criteria proposed in ref. 14 were tested for the inference of chromothripsis. Circos plot displays SNP ratio (inner circle, red dashed line indicating heterozygosity), CNA (outer circle, blue area indicating deletion, red amplification) and structural variations (colours as in v) as detected by WGS. Chr4 shows a complex deletion pattern and massive rearrangements associated with loss of one copy of *Cdkn2a*. The second copy of *Cdkn2a* is focally deleted. In addition, a balanced translocation of an approximately 200 kb segment from trisomic chr6 to chr4 and a far smaller segment of chr4 into chr6 was detected. The *Kras* locus is not directly affected by this inter-chromosomal translocation. LOH, CNAs and rearrangements are not detected on other chromosomes. q, In a chromothriptic model, DNA breakpoints tend to cluster on a chromosome. Testing against an exponential distribution (parameter λ derived from mean of observed distance between adjacent breakpoints) revealed significantly shorter distances than expected in a progressive model ($n = 146$ breakpoints). $P < 10^{-12}$; χ^2 goodness-of-fit

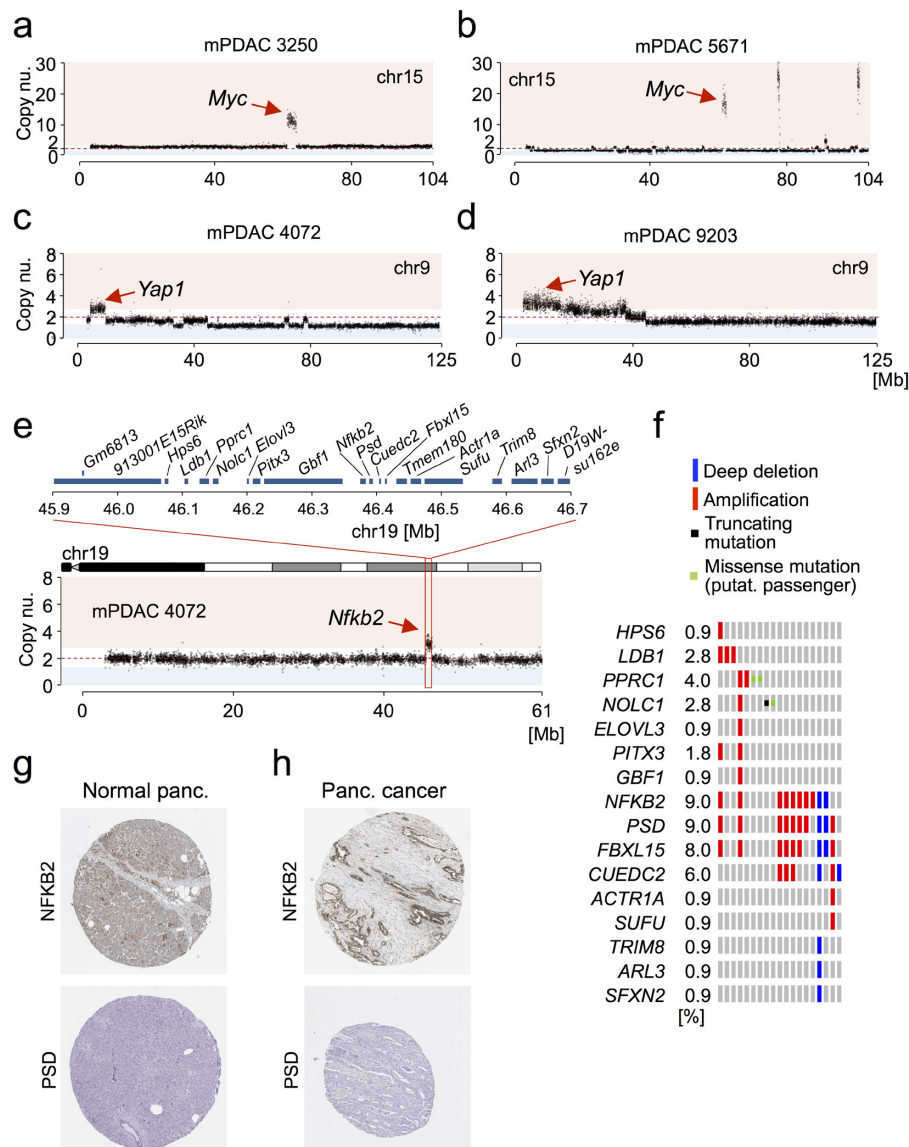
test. r, In a progressive model of acquisition of massive rearrangements or structural variations, copy-number states tend to be more complex than in the chromothriptic model. Monte Carlo simulations were used to generate a progressive evolution model with sequential accumulation of observed rearrangements ($n = 100$ simulations per number of structural variations). mPDAC S821 showed fewer copy-number states on chr4 than expected in the progressive model. Mean is indicated as a black point and lines represent the 95% confidence interval. s, Chromothriptic tumours typically feature interspersed loss and retention of heterozygosity. Accordingly, there was a high overlap between deleted regions and LOH segments on chr4 (Jaccard index (J) = 0.99). t, In a chromothriptic model, DNA shattering typically occurs on a single haplotype. M-FISH showed that significant loss of chromosomal content occurred on only one copy of chr4. u, To show random chromothriptic DNA shattering and re-joining, observed segments ($n = 73$) were re-ordered by running Monte Carlo simulations ($n = 10^3$) generating a background probability distribution. S821 segment order lies within the chromothriptic null model. Two-sided $P = 0.78$. v, All four structural variation types are uniformly distributed in a chromothriptic tumour model. $P = 0.43$; χ^2 goodness-of-fit test. w, In a chromothriptic model, paired-end connection types (as given by the structural variation type) induce an alternating sequence of DNA segment ends when ordered according to the genomic position on the original chromosome. Tendency towards this alternating 3'-to-5' pattern of rearranged DNA segment ends ($n = 146$) was tested by using right-sided Wald–Wolfowitz runs test. $P < 10^{-12}$. x, Mutation clusters in relation to breakpoint junctions involved in chromothripsis are shown as a rainfall plot for primary PDAC from PK mouse S821. Each dot represents a single SNV and is ordered on the x axis according to its position in the mouse genome. The distance of each SNV to the previous SNV in the genome is shown on the y axis. The colouring of individual SNV dots indicates the type of nucleotide substitution. y, Chr4 'zoom-in' from x. Breakpoint junctions are shown according to their genomic position on chr4. No mutation clusters—either in absence or in combination with breakpoint junctions—were detected, consistent with chromothripsis involving end joining DNA repair mechanisms. This is in contrast to other complex rearrangement types, such as chromoanagenesis, which arise through replication-based mechanisms with breakpoint-associated high mutation rates (for example, kataegis).



Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Specificity, timing, mechanisms and impact of *Kras*^{G12D} gene dosage alterations on gene expression in pancreatic tumorigenesis. **a**, Overlay of copy number profiles of primary mPDAC cell cultures from PK mice ($n = 38$) as determined by aCGH. The y axis shows the frequency of a genomic region to be amplified (up) or deleted (down) in the cohort, with *Cdkn2a* and *Kras* loci being most frequently affected by CNAs. **b**, Prevalence of LOH in primary mPDAC cell cultures from PK mice ($n = 38$) on the basis of WES data. A chromosome was considered to be affected by LOH if the SNP frequency was shifted to no more than 0.1 or at least 0.9 in a segment with a size of at least 200 kb. LOH on chr4 is frequently the consequence of heterozygous deletions involving the *Cdkn2a* locus. By contrast, LOH on chr6 is predominantly copy number neutral and linked to increased *Kras*^{G12D} gene dosage. Chr4 (home of *Cdkn2a*) and chr6 (home of *Kras*) show markedly increased rates of LOH compared with all other chromosomes, reflecting their functional importance during tumorigenesis. **c–h**, Genetic mechanisms of *Kras*^{G12D} gene dosage alterations as identified by aCGH, M-FISH and WES in pancreatic cancers from PK mice. The observed types of increased *Kras*^{G12D} gene dosage acquisition were (1) focal gain (affecting no more than 50% of the chromosome length), arising either through replication-based mechanisms (two cases, one with high-level *Kras*^{G12D} amplification (shown in **c**) and one with low level amplification) or translocation and subsequent amplification of the translocated chromosome (one case (shown in **d**)), (2) arm-level gain (affecting at least 50% of the chromosome length) arising through mitotic errors (seven cases of whole-chromosome gain (example shown in **e**), occasionally (two cases) with concomitant intra-chromosomal deletions or translocations not affecting *Kras* (example shown in **f**) and (3) copy-number-neutral LOH (CN-LOH, *Kras*^{G12D} homozygosity, acquired uniparental disomy), arising either through mitotic recombination (affecting parts of chr6 (shown in **h**)) or chromosomal missegregation (duplication of *Kras*^{G12D}-mutant chr6 and loss of wild-type chr6 (shown in **g**)). **c**, mPDAC S134 shows a high-order focal amplification of *Kras*^{G12D}. The sharp borders, small size of the amplification (600 kb) and strong increase in copy number ($4\times$) indicate that *Kras*^{G12D} was amplified through multiple cycles of repeated template-switching by a replication-based DNA repair mechanism. The *Kras*^{G12D} mutant allele frequency is 89.1%. **d**, Tumour 4706 carries a focal amplification of *Kras*^{G12D}. M-FISH analysis revealed that the mutant *Kras*^{G12D} allele (chr6) was probably first affected by a reciprocal translocation of chr4 and chr6, resulting in two rearranged chromosomes: der(4)T(4;6) and der(6)T(4;6). Subsequently, der(4)T(4;6) was missegregated through mitotic error, resulting in focal gain of the *Kras*^{G12D} locus. The *Kras*^{G12D} mutant allele frequency is 72.2%. **e**, mPDAC R1035

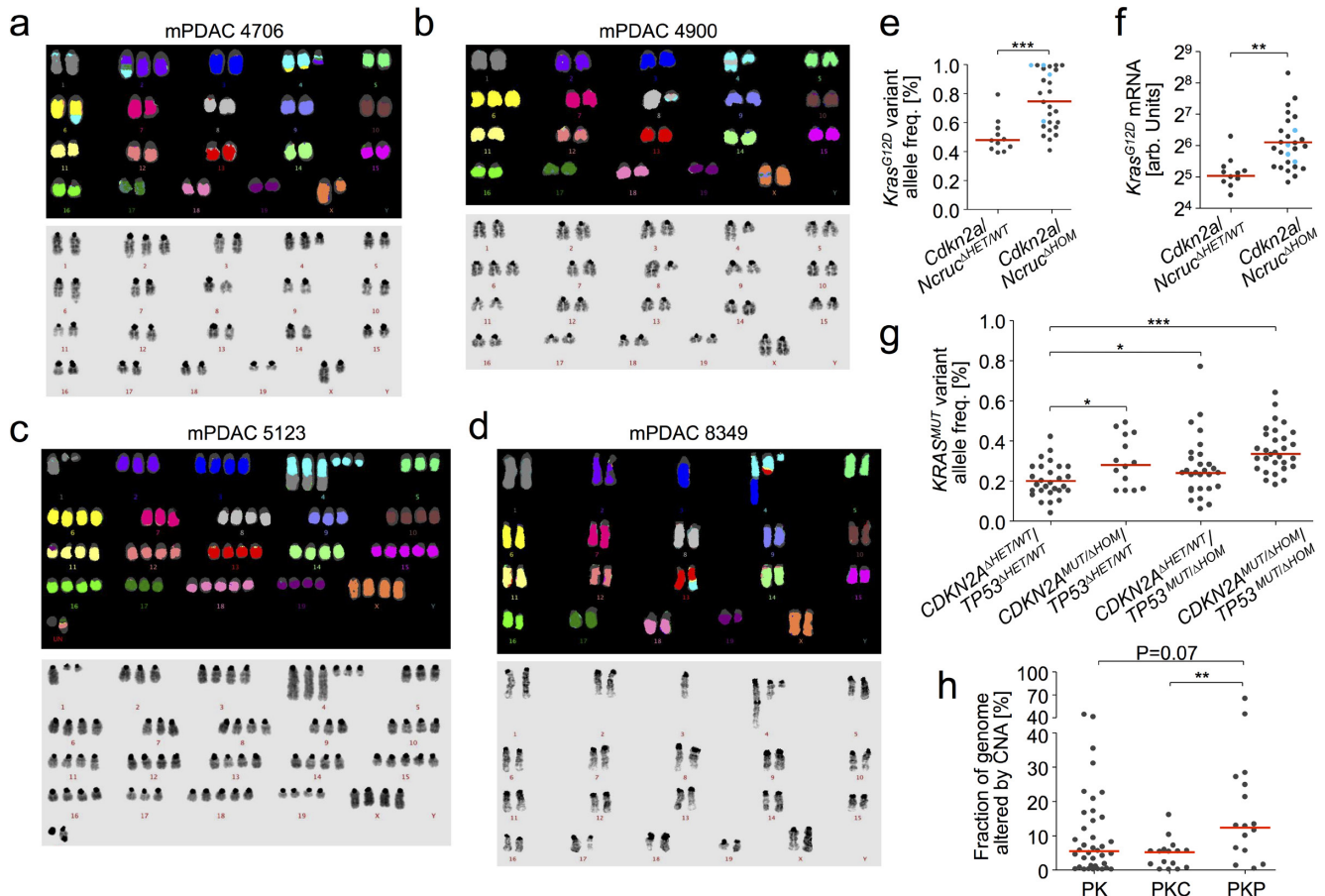
shows 'classic' whole-chromosome gain (trisomy) of chr6, which was probably generated through mitotic error or missegregation. The *Kras*^{G12D} mutant allele frequency is 69.8%. **f**, In tumour 8442, arm-level gain of *Kras*^{G12D} was probably generated through mitotic missegregation of chr6. Intra-chromosomal deletion on one of three chromosomes (19.6 Mb) does not affect *Kras*. The *Kras*^{G12D} mutant allele frequency is 66.4%. Asterisk, chr6 with reduced length resulting from intra-chromosomal deletion. **g, h**, mPDAC 16992 and B590 display CN-LOH, leading to increased *Kras*^{G12D} gene dosage. The *Kras*^{G12D} mutant allele frequencies are 99.2% and 96.3%, respectively. The SNP pattern of chr6 in mPDAC 16992 reveals that the whole chromosome is affected by CN-LOH, indicating chromosome missegregation (duplication of the *Kras*^{G12D}-mutant chr6 and loss of wild-type chr6) as the underlying mechanism. By contrast, in mPDAC B590 only a partial region of chr6 is affected by CN-LOH, therefore probably resulting from mitotic recombination. **i**, Allele-specific *Kras*^{G12D} mRNA expression in *Kras*^{G12D-HET} ($n = 12$) versus *Kras*^{G12D-IGD} ($n = 26$) primary PDAC cell cultures from PK mice as detected by combined analysis of amplicon-based RNA-seq (proportion of mutant/wild-type *Kras* mRNA) and 3' polyadenylation RNA-seq (amount of total *Kras* mRNA, but not the proportion of mutant/wild-type *Kras* mRNA, owing to sequencing of 3' transcript ends; see Methods). This figure is related to Fig. 2b. *** $P \leq 0.001$, two-tailed Mann–Whitney test; bars, median. **j**, Mutant *Kras*^{G12D} mRNA levels in *Cdkn2a/Ncruc*^{ΔHET/WT} ($n = 11$) versus *Cdkn2a/Ncruc*^{ΔHOM} ($n = 27$) primary PDAC cell cultures from PK mice as detected by combined amplicon-based RNA-seq and 3' polyadenylation RNA-seq. This figure is related to Extended Data Fig. 5f. *** $P \leq 0.001$, two-tailed Mann–Whitney test; bars, median. **k**, Mutant *Kras*^{G12D} mRNA levels in transcriptional clusters of mPDAC from PK mice (C2a/b/c/C1, $n = 5/7/6/15$) as detected by combined amplicon-based RNA-seq and 3' polyadenylation RNA-seq. This figure is related to Fig. 5d. $P = 1.6 \times 10^{-5}$, two-sided Pearson correlation; bars, median. **l–n**, Interphase FISH for the analysis of copy-number and ploidy states at the KRAS locus on chr12 in hPanIN with KRAS^{G12} variant allele frequencies (VAFs) of approximately 100%. KRAS^{G12} VAFs are indicated above each FISH profile as detected by amplicon-based deep sequencing. A VAF of approximately 100% can be caused either by loss of the wild-type KRAS locus (hemizygosity of KRAS^{G12-MUT}; one KRAS^{G12-MUT} allele per cell) or by CN-LOH (acquired uniparental disomy; homozygosity of KRAS^{G12-MUT}; two KRAS^{G12-MUT} alleles per cell). All samples show a diploid genome as suggested by CEN12 (centromere probe chr12; two red signals per nucleus). Neither loss of one KRAS allele nor monosomy of chr12 was observed, providing evidence for CN-LOH and increased KRAS^{G12-MUT} gene dosage in hPanIN. Scale bars, 2.5 μm.



Extended Data Figure 4 | Enrichment for amplification of alternative oncogenic drivers in mPDACs of PK mice with *Kras*^{G12D-HET} status.

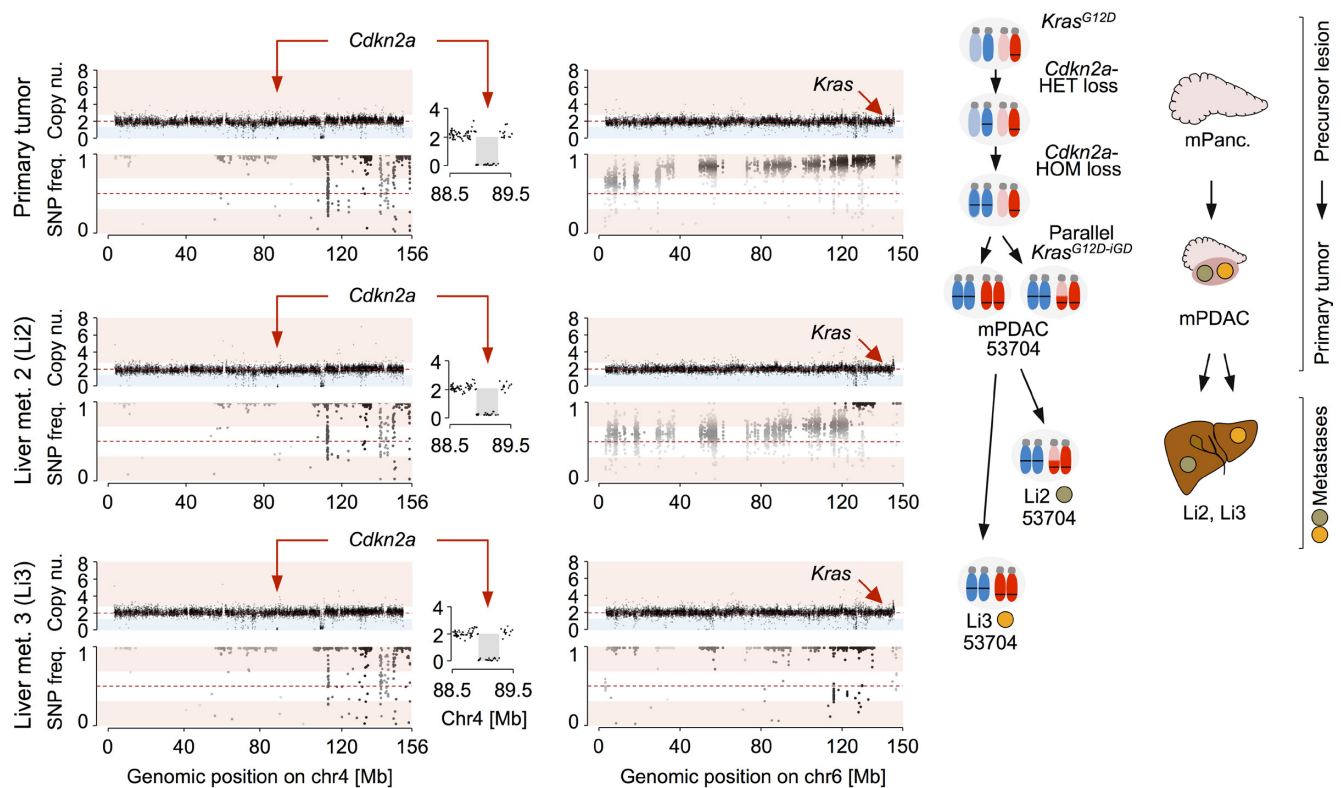
a, b, Two primary mPDACs with strong focal *Myc* amplification on chr15 are shown, as detected by aCGH. Red dashed line indicates no copy-number change. **c, d**, Focal copy number gains targeting the *Yap1* locus on chr9 in primary mPDACs 4072 and 9203 as revealed by aCGH. **e**, Chr19 was also frequently subject to arm-level gain (see Fig. 1c and Extended Data Fig. 11). Primary mPDAC of PK mouse 4072 harbours a focal gain on chr19 containing 20 genes: *913001E15Rik*, *Gm6813*, *Hps6*, *Ldb1*, *Pprc1*, *Nolc1*, *Elovl3*, *Pitx3*, *Gbf1*, *Nfkb2*, *Psd*, *Fbxl15*, *Cuedc2*, *Tmem180*, *Actr1a*, *Sufu*, *Trim8*, *Arl3*, *Sfxn2* and *D19Wsu162e*. **f**, Cross-species analyses revealed that the orthologous region on human chr10 is also subject to

recurrent amplifications in human PDAC (8 out of 109 hPDACs have focal amplifications; data from ref. 6). Of the 20 mouse genes, 16 could be assigned to orthologues in humans. Further analyses revealed that only two genes, *NFKB2* and *PSD*, are within the minimal overlapping region of recurrent amplification (data from ref. 6 and oncoplot from cBioPortal^{60,61}). **g**, *NFKB2*, but not *PSD*, shows medium protein expression in exocrine glandular cells of normal pancreatic tissue, as detected by immunohistochemistry (data from the Human Protein Atlas⁶²). **h**, *NFKB2* is highly expressed in 17% (2 out of 12) of stained hPDAC biopsies as shown by immunohistochemistry. In contrast, there was no *PSD* expression in any of the analysed pancreatic cancers (0 out of 12). Protein expression data were used from the Human Protein Atlas⁶².



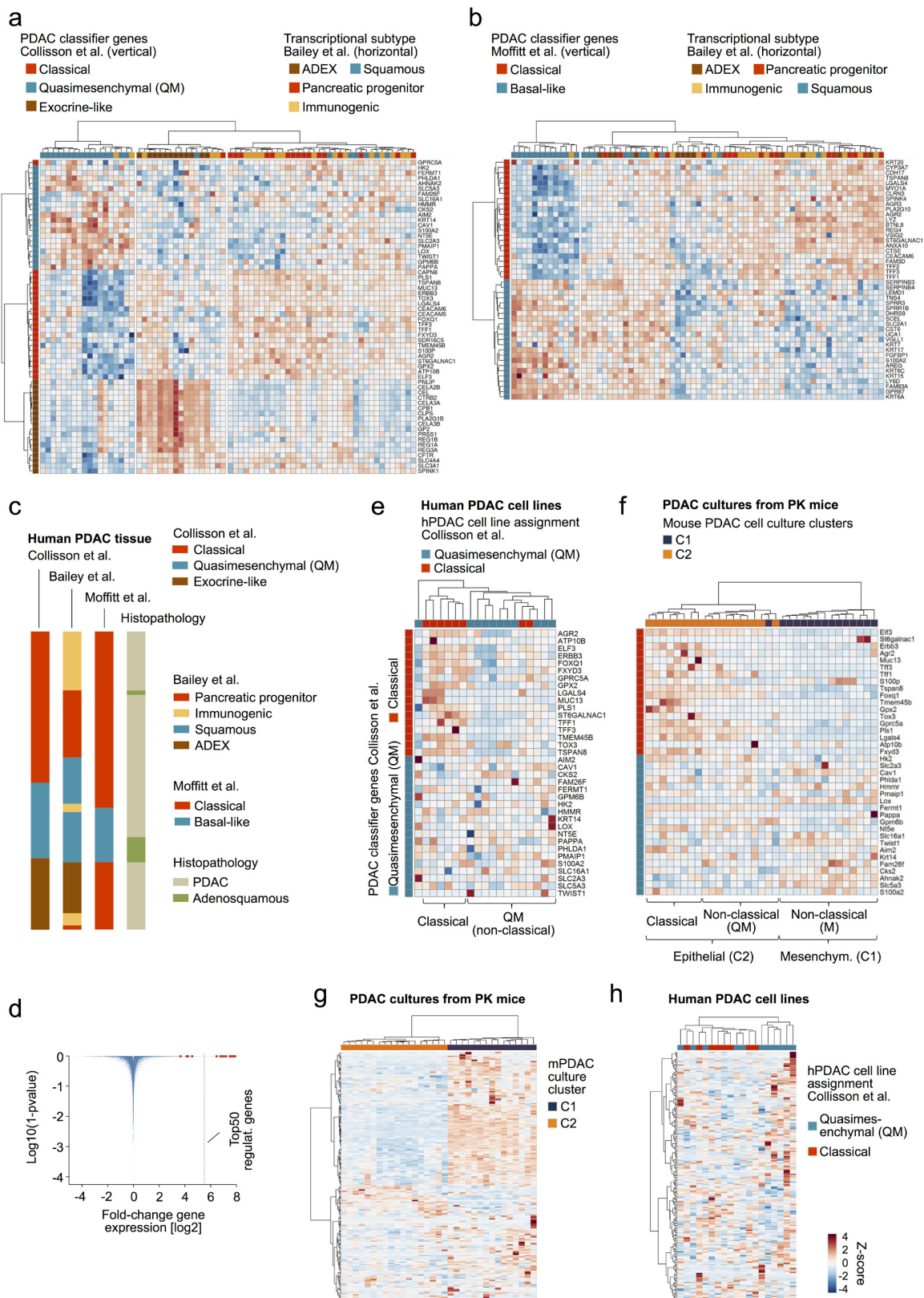
Extended Data Figure 5 | Characterization of *Cdkn2a* (chr4) alterations and correlation with *Kras*^{MUT} gene dosage variation and mRNA expression in mouse and human PDAC. a–d, *Cdkn2a* alteration on mouse chr4 can occur through arm-level, complex or focal loss as well as uniparental disomy (see Fig. 3). In addition, chr4 is frequently involved in inter-chromosomal translocations. Examples of representative karyotypes of primary pancreatic cancer cultures derived from PK mice with translocations involving chr4, probably affecting the *Cdkn2a* locus. In all four cases, chr4 translocations were found in all ten metaphase spreads of each cancer, indicating their early acquisition during tumour evolution. **a**, mPDAC 4706 with diploid karyotype: 42, XX, del(X), +2, der(2)t(2;4)is(2;4), der(4)t(4;6)*2, +der(4)t(2;4), der(6)t(4;6). **b**, mPDAC 4900 also features a diploid karyotype: 41, XX, der(X)is(X;4), der(4)is(4;8), del(4), +6, der(8)t(4;8). **c**, mPDAC 5123 underwent polyploidization, after translocation of chr4 with chr1 and a deletion on the other copy: 78, XXXX, -1, del(1)*2, -2, +4*2, der(4)t(1;4)*3, del(4)*3, -5, -7, -9, +15, -17, +18. **d**, mPDAC 8349 shows a diploid karyotype: 40, XX, der(4)t(3;4), der(4)t(4;13), +del(4), der(13)t(4;13). **e**, *Kras*^{G12D} variant allele frequencies detected by amplicon-based deep sequencing of the *Kras* locus are higher in *Cdkn2a*/*Ncruc*^{ΔHOM} mPDAC (*n* = 27) compared with *Cdkn2a*/*Ncruc*^{ΔHET/WT} (*n* = 11) pancreatic cancers. All cancers are from PK mice. Blue dots indicate tumours with complete *Ncruc* deletion. ****P* ≤ 0.001, two-tailed Mann-Whitney test; bars, median. **f**, Allele-specific expression of mutant *Kras*^{G12D} mRNA is increased in primary

tumours from PK mice with *Cdkn2a*/*Ncruc*^{ΔHOM} (*n* = 27) background compared with *Cdkn2a*/*Ncruc*^{ΔHET/WT} (*n* = 11) cancers. Primary mPDACs with homozygous loss of *Ncruc* are highlighted in blue. *Kras*^{G12D} expression was analysed by combining amplicon-based RNA-seq and qRT-PCR (as described in the Methods section). ***P* = 0.003, two-tailed Mann-Whitney test; bars, median. **g**, *KRAS*^{MUT} variant allele frequencies based on WES in a published data set of microdissected human PDAC (ref. 6, reduced stromal content) was analysed with respect to *CDKN2A* and *TP53* status. *KRAS*^{MUT} allele frequency was higher in mutated/homozygously deleted *CDKN2A* and/or *TP53* (*CDKN2A*^{MUT/ΔHOM}/*TP53*^{MUT/ΔHOM}; hPDACs compared with cancers with *CDKN2A*^{ΔHET/WT}/*TP53*^{ΔHET/WT} status (from left: *n* = 28, *n* = 14, *n* = 28, *n* = 30). Two-sided rank-based ANOVA (*P* = 5.8 × 10⁻⁶); post hoc testing with two-sided Tukey honest significant difference test, *adjusted *P* ≤ 0.05, ***adjusted *P* ≤ 0.001; bars, median. **h**, Fraction of the genome altered by copy number changes detected by aCGH in primary mPDACs of PK (*n* = 38), PKC (*n* = 16) and PKP (*n* = 16) mice. PKP mice show a significantly increased CNA load compared with PKC mice. Two-sided rank-based ANOVA (*P* = 0.01); post hoc testing with two-sided Tukey honest significant difference test, **adjusted *P* = 0.009, adjusted *P* values for group-wise comparisons are shown; bars, median. Del, deletion; der, derivative chromosome; is, insertion; t, translocation; '−', chromosome loss; '+', chromosome gain.



Extended Data Figure 6 | Complete *Cdkn2a* barrier loss precedes *Kras*^{G12D-IGD} in primary mPDAC of PK mouse 53704. CNAs at chr4 (*Cdkn2a*) and chr6 (*Kras*) in mPDAC 53704 and corresponding metastases, as detected by aCGH (top) and WES-based SNP pattern analysis (bottom). The primary cancer and both liver metastases display identical focal deletions of *Cdkn2a* and similar SNP patterns on chr4, revealing that all lesions share the same ancestor cell with complete *Cdkn2a* loss. By contrast, SNP analysis on chr6 revealed discordant patterns in the primary mPDAC and both metastases. Li2 shows partial LOH of a distal region on chr6 involving the *Kras* locus, while LOH in Li3 involves the whole chr6. This explains the stepwise LOH pattern observed on chr6 in the primary mPDAC. The graphic on the right shows the combined interpretation of CNA and LOH profiles, which suggests the following sequence of genetic events during tumour evolution. The initial *Kras*^{G12D} mutation was followed by focal deletion of one copy of *Cdkn2a*. In a subsequent genetic event, the second copy of *Cdkn2a* was

lost by chr4 missegregation and copy-number-neutral LOH. Complete barrier loss allowed for convergent evolution of increased *Kras*^{G12D} gene dosage through copy-number-neutral LOH and gave rise to independent metastases in the liver. Note that a major obstacle for equivalent human studies is the limited availability of human matched primary-metastasis samples, particularly of treatment-naïve ones. We performed cross-species analyses using data from a recent study, which analysed human treatment-naïve metastatic PDACs by WGS⁸ and provided *CDKN2A* and *KRAS* copy number data for matched primaries and metastases from three patients. In one patient the sequential order of *CDKN2A* deletion and *KRAS* amplification could be reconstructed: homozygous *CDKN2A* deletions were identical in all primaries and metastases, whereas there were five different *KRAS* gains in the six metastases. This suggests convergent evolution of mutant *KRAS* gene dosage gain upon homozygous *CDKN2A* loss in this patient, in line with similar data in large series of mouse cancers and their metastases (see Fig. 3e).

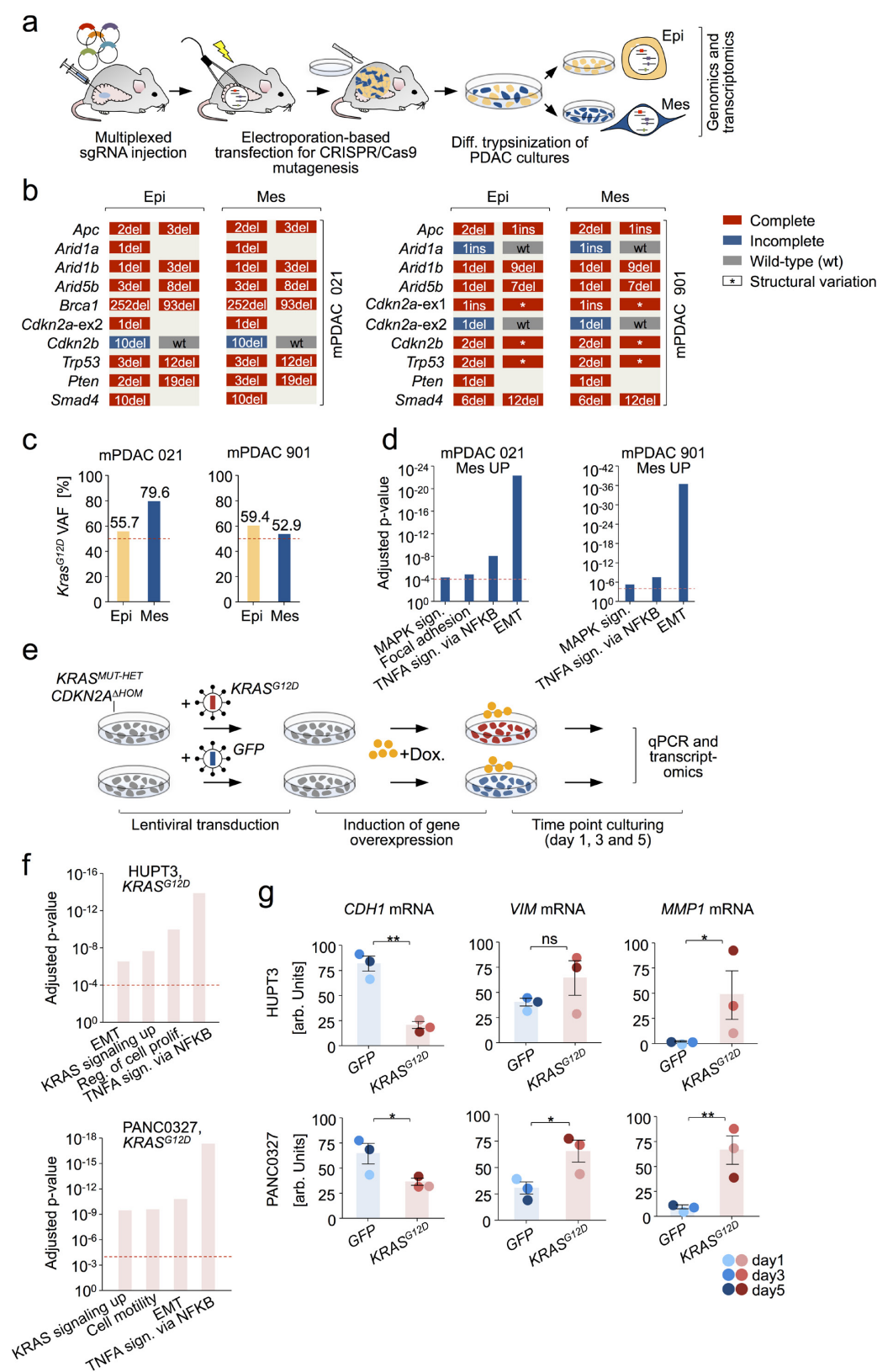


Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Transcriptome-based subtyping of human primary pancreatic cancer and classification of human PDAC cell lines and primary PDAC cell cultures from PK mice. a–c, Independent cross-comparison of transcriptional classification systems from Collisson *et al.*²⁸, Moffitt *et al.*²⁹ and Bailey *et al.*⁷. Collisson *et al.*²⁸ performed PDAC microdissection and defined three transcriptional subtypes: classical, quasimesenchymal and exocrine-like. Moffitt *et al.*²⁹ defined two subtypes (classical, basal-like) using (1) virtual separation of tumour and non-tumour gene expression patterns, (2) transplantation studies and (3) human PDAC cell lines; and they proposed that the exocrine-like signature stems from exocrine pancreatic cells, rather than from the cancer cells. Bailey *et al.*⁷ used bulk tumours and defined four subtypes (pancreatic progenitor, immunogenic, squamous, aberrantly differentiated endocrine exocrine (ADEX)). RNA-seq data from PDAC and adenosquamous pancreatic carcinoma from Bailey *et al.*⁷ were used for cross-comparison of classification systems. Other histological subentities of pancreatic cancer were excluded (for example, intraductal papillary mucinous neoplasm, mucinous cystic neoplasm, acinar cell carcinoma). The subtyping from Bailey *et al.*⁷ for this data set was available.

a, Unbiased hierarchical clustering of primary pancreatic cancer samples ($n = 71$) from Bailey *et al.*⁷ using classifier genes of Collisson *et al.*²⁸. **b**, Subtyping of primary pancreatic cancer samples ($n = 71$) from Bailey *et al.*⁷ using classifier genes defined by Moffitt *et al.*²⁹. **c**, Consensus clustering based on analyses performed in **a** and **b**. There is considerable overlap between at least two subtypes, which are in large parts captured by the initially proposed classical and quasimesenchymal signatures from Collisson *et al.*²⁸ (which are also detected in mouse and human PDAC cell lines; see Extended Data Fig. 7e–h). The classification of Bailey *et al.*⁷ (based on bulk tissue analyses) suggests that classical cancers of Collisson *et al.*²⁸ (microdissected cancer tissue) can be further sub-stratified into some with and some without a strong immune cell infiltration. The classification of Moffitt *et al.*²⁹ suggests that the exocrine-like signature of Collisson *et al.*²⁸ (ADEX subtype in Bailey *et al.*⁷) stems from ‘contaminating’ healthy exocrine pancreatic cells, on the basis of the evidence described above. Given that the exocrine-like signature of Collisson *et al.*²⁸ was derived from microdissected PDAC, such ‘contamination’ is only conceivable if exocrine-like signature genes

were much more highly expressed in pancreatic acinar cells than in PDAC cells. **d**, Volcano plot showing strongly upregulated expression of exocrine-like genes in human wild-type pancreas (13- to 241-fold; median 183-fold upregulation). Note that 15 out of 19 exocrine-like signature genes (red dots) are among the top 50 genes upregulated in human wild-type pancreas ($n = 3$) compared with hPDAC cell lines ($n = 30$) (y axis is calculated on Benjamini–Hochberg adjusted P values derived from R package limma (see Methods)). Although these data do not exclude the existence of exocrine-like PDACs, they support the possibility that ‘contamination’ with few acinar cells can impose an exocrine-like signature on a cancer. This might explain why human or mouse PDAC cell lines do not cluster into the exocrine-like subtype (see also Extended Data Fig. 7e, f). **e**, Hierarchical clustering of microarray-based expression profiles using identifier genes from Collisson *et al.*²⁸ on human PDAC cell lines ($n = 19$, Gene Expression Omnibus series GSE17891). As also described in Collisson *et al.*²⁸, only two subtypes can be detected in human cell line collections: classical and quasimesenchymal. Of note, the most prominent change in the quasimesenchymal cell lines is downregulation (extinction) of the classical assigner genes, whereas expression of quasimesenchymal classifier genes is quite variable. We therefore also use here the terms ‘classical’ and ‘non-classical’. **f**, Projection of the Collisson *et al.*²⁸ classifiers on to mouse PDAC cell culture transcriptomes ($n = 33$) also identified classical and non-classical subtypes. The non-classical subtype contained a subset of mPDAC cell cultures from cluster C2a/b/c (epithelial morphology; equivalent of human quasimesenchymal) and all cluster C1 mPDACs (mesenchymal morphology; ‘M’ cluster). **g**, Application of a human EMT hallmark gene set⁵² for hierarchical clustering of expression profiles from primary PDAC cultures (PK mice; $n = 33$) resulted in a separation of C1 (mesenchymal) and C2a/b/c (epithelial) cell lines. **h**, Projection of the EMT hallmark gene set on human PDAC cell line transcriptomes ($n = 19$, Gene Expression Omnibus series GSE17891) did not result in a clear separation of samples, indicating underrepresentation of the mesenchymal M subtype (equivalent to mouse C1/‘M’) in available human cell line collections. As shown in Extended Data Fig. 9b, however, the EMT signature is detectable in undifferentiated human pancreatic carcinoma, which is the human equivalent of the mesenchymal mouse PDACs in C1.

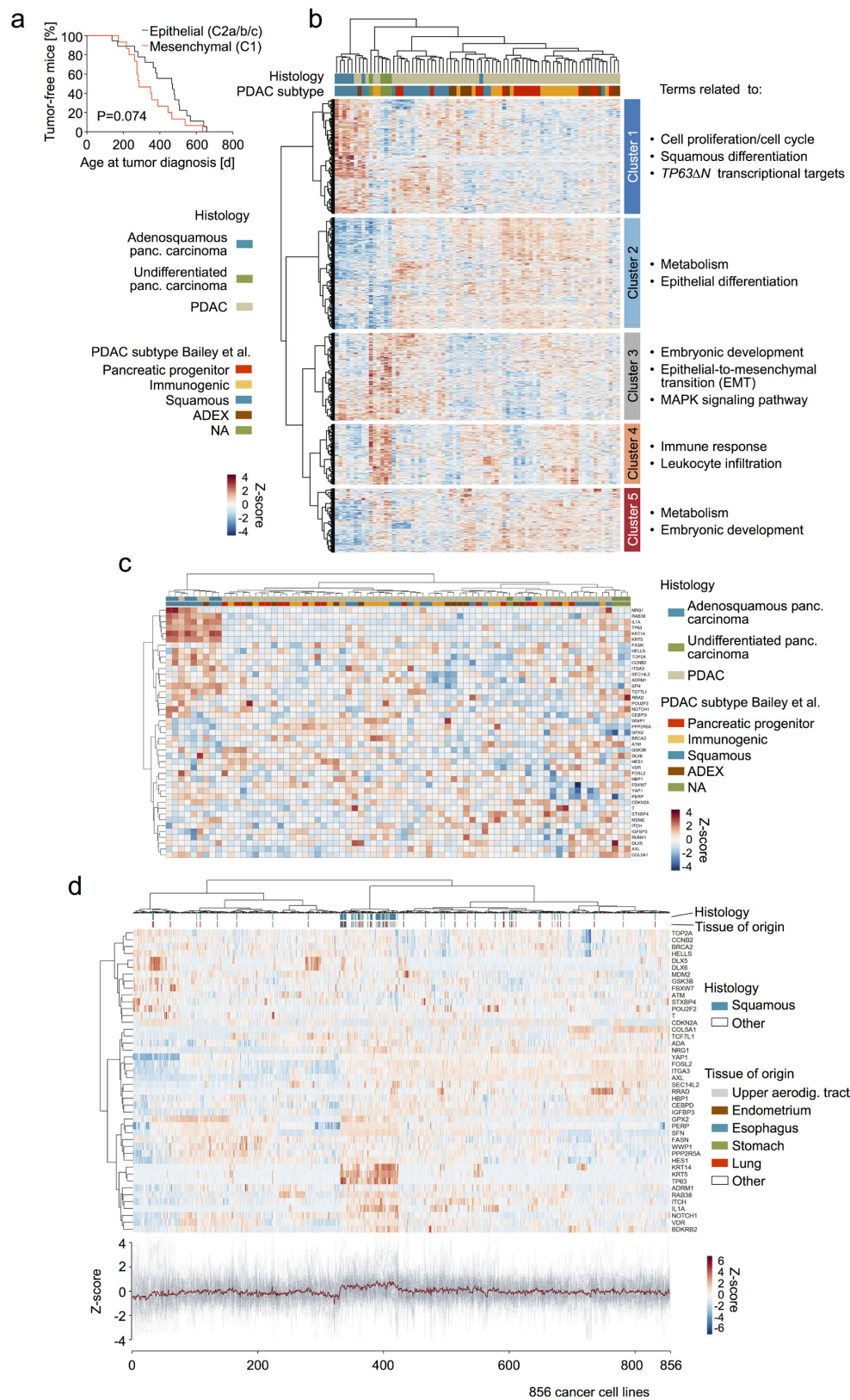


Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Functional analyses to study the role of increased *Kras*^{G12D} gene dosage in EMT. a–d, Multiplexed somatic CRISPR–Cas9 mutagenesis for phylogenetic tracking of epithelial–mesenchymal mPDAC clones *in vivo*.

a, Major steps of multiplexed gene editing by pooled delivery of CRISPR–Cas9 vectors, each targeting a different tumour suppressor gene in the pancreas of PK mice. Electroporation-based transfection enables low-frequency mosaic vector delivery (average of 120 cells per pancreas are transfected) to induce clonal tumours. Primary tumour cell cultures were screened for the simultaneous presence of epithelial and mesenchymal cells. Two such cancers were identified (mPDACs from mouse 021 and mouse 901) and subjected to differential trypsinization to enrich for each morphology. **b**, Amplicon-based deep sequencing of all single-guide RNA-targeted loci revealed identical indel patterns in both epithelial–mesenchymal culture pairs. This shows (1) that epithelial and mesenchymal cells originate from the same clone and (2) that the CRISPR-induced mutations do not contribute to the differential phenotype. **c**, *Kras*^{G12D} VAFs in epithelial and mesenchymal cell cultures from mPDAC 021 and mPDAC 901, as detected by amplicon-based deep sequencing. Both cancers had increased *Kras*^{G12D} expression in mesenchymal cells (see Fig. 5e). In mPDAC 021, this is due to selective amplification of the *Kras*^{G12D} allele in mesenchymal cells. In mPDAC 901, genetic *Kras*^{G12D} amplification was not observed, suggesting induction of increased *Kras* expression in mesenchymal cells by other mechanisms. **d**, Gene set enrichment analysis using MSigDB of differentially regulated genes in mesenchymal versus epithelial mPDACs on the basis of RNA-seq. Mesenchymal clones of mPDAC 021 and mPDAC 901 show an upregulation of genes involved in

‘MAPK signaling pathway’ and ‘EMT’ compared with the corresponding epithelial clones, in line with increased *Kras*^{G12D} gene dosage (a full list of enriched gene sets is provided for comparison in Supplementary Table 15). False discovery rate-adjusted *P* values are shown on the *y* axis. Representative data from one experiment are shown. **e–g**, Induction of EMT-like transcriptional programs by *KRAS*^{G12D} overexpression in human PDAC cell lines. **e**, Graphic of experimental workflow. Two human PDAC cell lines (HUPT3 and PANC0327) with homozygous *CKDN2A* loss (*CKDN2A*^{ΔHOM}) and heterozygous *KRAS*^{MUT} (*KRAS*^{MUT-HET}) status were transduced with lentivirus carrying doxycycline-inducible *KRAS*^{G12D} or GFP-control expression constructs. *KRAS*^{G12D} or GFP expression was induced by adding doxycycline for 1, 3 or 5 days. **f**, Gene set enrichment analysis using MSigDB of differentially regulated genes in *KRAS*^{G12D}- versus GFP-induced hPDAC cell lines HUPT3 and PANC0327 on the basis of RNA-seq. Upon doxycycline treatment, both hPDAC cell lines showed consistent upregulation of genes involved in ‘KRAS signaling up’ and ‘EMT’ (a full list of enriched gene sets is provided for both cell lines in Supplementary Table 16). False discovery rate-adjusted *P* values are shown on the *y* axis. **g**, Expression of marker genes for epithelial (*CDH1*) or mesenchymal (*VIM*) cell differentiation and invasion or matrix disassembly (*MMP1*) was validated by qPCR (normalized to *GAPDH* and *PPIA*). In line with RNA-seq data, *KRAS*^{G12D}-induced cells show an increased expression of the mesenchymal marker gene *VIM*, increased expression of *MMP1* and reduced levels of epithelial marker gene *CDH1*. **P* ≤ 0.05, ***P* ≤ 0.005, NS, not significant, two-tailed *t*-test; bars, mean; error bars, s.e.m.

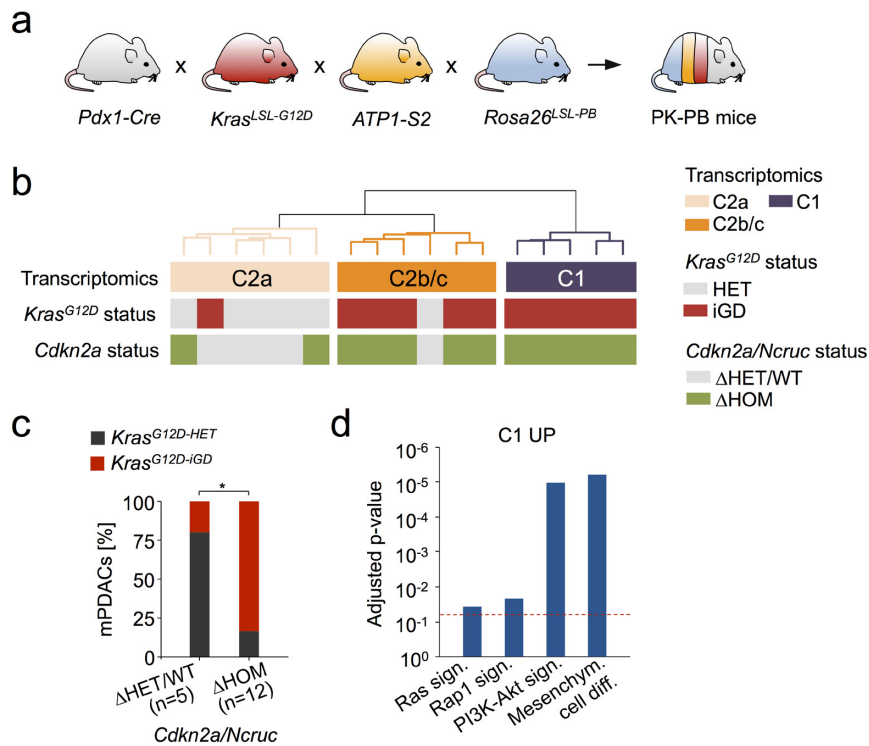


Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Transcriptional profiles of human undifferentiated pancreatic carcinomas are enriched for signatures of oncogenic signalling intensification and EMT but not for activation of *TP63ΔN* transcriptional network.

a. Primary pancreatic tumours from PK mice with a mesenchymal phenotype (C1 cluster, $n = 15$) are almost exclusively classified as undifferentiated or sarcomatoid by histopathological evaluation and tend to have a reduced age at diagnosis when compared with epithelial (C2a/b/c cluster, $n = 18$) tumours (histopathological grade 1–3 (G1–G3)). This aggressive behaviour of undifferentiated pancreatic carcinoma is also observed in human patients and is associated with worse clinical outcome³³. P value calculated by two-sided log-rank test. **b.** Comparison of publicly available expression profiles of human undifferentiated pancreatic carcinoma ($n = 4$), PDAC (World Health Organization grades 1–3 (G1–G3), $n = 64$) and adenosquamous pancreatic carcinoma ($n = 7$). Human samples with the above histopathological characteristics for which expression-based subtype information from Bailey *et al.*⁷ was available were used and complemented with available undifferentiated pancreatic carcinomas from the ICGC PACA-AU cohort (Supplementary Table 18). Other histological subentities of pancreatic cancer were excluded (for example, intraductal papillary mucinous neoplasm, mucinous cystic neoplasm, acinar cell carcinoma). ANOVA was performed to select genes that were differentially expressed in at least one of the six defined subgroups of pancreatic cancer: (1) undifferentiated, (2) adenosquamous pancreatic carcinoma and (3–6) PDAC (G1–G3) sub-stratified in pancreatic progenitor, immunogenic, squamous and ADEX subtypes from Bailey *et al.*⁷. Differentially regulated genes were used for unbiased hierarchical clustering of these pancreatic cancer transcriptional profiles. Five sub-clusters of co-regulated gene expression could be identified according to the cluster tree on the y axis (separated by white horizontal bars in the heatmap). Gene set enrichment analysis using MSigDB was performed for individual sub-clusters and terms related to predominating gene sets or pathways are annotated for each cluster on the right (full list provided in Supplementary Table 17). Undifferentiated pancreatic carcinomas cluster together and are associated with (1) upregulation of genes in cluster 3 (containing MAPK signalling pathway and gene sets relevant during embryonic development or EMT) and (2) downregulation of

genes in clusters 2 and 5, which contain gene sets related to epithelial cell differentiation, embryonic development or metabolic signatures. This reflects the pathway enrichment signature in the equivalent undifferentiated (mesenchymal) mouse PDACs (cluster C1/'M' in PK mice; see Extended Data Fig. 7g) and provides further support for the link between *KRAS* signalling intensification, EMT and the undifferentiated tumour phenotype. The immunogenic PDAC subtype showed high expression of cluster 4 genes, which was also strong (even elevated) in undifferentiated pancreatic carcinomas, suggesting increased immune cell infiltration in undifferentiated carcinomas. Cluster 1 contained gene sets related to cell proliferation/cell cycle, squamous differentiation and *TP63ΔN* transcriptional targets, which were most highly overexpressed in pancreatic carcinomas with adenosquamous histology. Undifferentiated pancreatic carcinomas did not show activation of the *TP63ΔN* transcriptional targets. This suggests that activation of *TP63ΔN* transcriptional targets is not causally linked to *KRAS* signalling intensification and EMT (see also Extended Data Fig. 9c, d, showing a lack of association of undifferentiated carcinomas with *TP63ΔN* transcriptional network activation). **c.** Unbiased hierarchical clustering of human pancreatic carcinomas with adenosquamous histology ($n = 7$) as well as PDACs (World Health Organization grades 1–3 (G1–G3), $n = 64$) and undifferentiated pancreatic carcinomas ($n = 4$) (sample set as in Extended Data Fig. 9b) using a list of validated *TP63ΔN* transcriptional targets⁵³. Pancreatic cancers with adenosquamous differentiation were significantly enriched in a cluster showing increased *TP63ΔN* transcriptional network activity ($P \leq 0.001$, two-sided Fisher's exact test, odds ratio 130, 95% confidence interval 11.6–1,452). Undifferentiated pancreatic carcinomas did not contribute to this cluster. In line with these results, pancreatic cancers from PK mice did not show differential regulation of the *TP63ΔN* network, reflecting the lack of adenosquamous tumours in this cohort (not shown). **d.** Unbiased hierarchical clustering across solid cancers (Cancer Cell Line Encyclopedia, $n = 856$) using the same gene list showed a strong enrichment of tumours with squamous differentiation in the sub-cluster with the highest *TP63ΔN* transcriptional network expression ($P \leq 0.001$, two-sided Fisher's exact test, odds ratio 28.1, 95% confidence interval 16.4–48.1), in line with the observation in ref. 63 that *TP63ΔN* is a signature for squamous differentiation across cancers.



Extended Data Figure 10 | *Kras^{G12D}* gene dosage is a critical determinant of PDAC biology in a mouse model with high mutational load. **a**, The mutational burden in primary PDAC cultures of PK mice was significantly lower compared with human PDAC studies (see Fig. 1b). To account for this potential confounding factor and to test whether our discoveries in PK mice also applied in a setting of high mutational burden, we used a mouse model combining *Kras^{G12D}* mutation and *PiggyBac* transposon-based insertional mutagenesis (PK–PB mice¹³). PK–PB mice show accelerated tumorigenesis compared with PK mice. PK–PB-derived tumours had an extensive mutational burden (median of 494 transposon insertions per tumour). Primary cultures of PDAC from PK–PB mice ($n=17$) were subjected to comprehensive genetic characterization using aCGH, microarray-based gene expression profiling, quantitative transposon insertion-site sequencing and amplicon-based deep sequencing of the *Kras* locus. **b**, Transcriptome profiles of primary PDAC cultures from PK–PB mice ($n=17$) were used for unbiased hierarchical clustering that resulted in two major clusters (C1 and C2), as in PK mice. *Kras^{G12D}* gene dosage status (as determined by aCGH and amplicon-based deep sequencing of the *Kras* locus) and *Cdkn2a* status (as determined by aCGH and quantitative transposon insertion-site sequencing) are indicated below the cluster tree for each individual

tumour. Similarly to PK mice, cluster C2a was characterized by *Kras^{G12D-HET}* and *Cdkn2a/Ncruc^{\Delta}HET/WT* status, whereas mPDACs in clusters C2b/c and C1 had increased *Kras^{G12D}* gene dosage (*Kras^{G12D-iGD}*) and were *Cdkn2a/Ncruc^{\Delta}HOM*. The genetic *Kras^{G12D}* status was significantly associated with expression clusters ($P=0.01$, two-sided Fisher's exact test), providing further evidence that expression clusters are associated with *Kras^{G12D}* gene dosage. **c**, Prevalence of *Kras^{G12D-iGD}* in cultures of primary mPDAC (from PK–PB mice) with homozygous ($n=12$) or heterozygous/wild-type ($n=5$) *Cdkn2a/Ncruc* status. $*P=0.03$, two-sided Fisher's exact test, odds ratio 20.0, 95% confidence interval 1.4–287.8. **d**, Gene set enrichment analysis using DAVID of upregulated genes in cluster C1 ($n=5$) compared with cluster C2 ($n=12$) of primary mPDAC cultures from PK–PB mice. As in PK mice, PK–PB tumours in C1 are characterized by upregulation of genes enriched in gene sets describing mesenchymal cell differentiation and revealed a strong enrichment for Ras downstream signalling pathways (full list in Supplementary Table 19). False discovery rate-adjusted P values are shown on the y axis. Overall, these analyses show that the biological principles discovered in the PK model also apply to pancreatic cancers from PK–PB mice with high mutational load.

Strong disk winds traced throughout outbursts in black-hole X-ray binaries

B. E. Tetarenko¹, J.-P. Lasota^{2,3}, C. O. Heinke¹, G. Dubus⁴ & G. R. Sivakoff¹

Recurring outbursts associated with matter flowing onto compact stellar remnants (such as black holes, neutron stars and white dwarfs) in close binary systems provide a way of constraining the poorly understood accretion process. The light curves of these outbursts are shaped by the efficiency of angular-momentum (and thus mass) transport in the accretion disks, which has traditionally been encoded in a viscosity parameter, α . Numerical simulations^{1–3} of the magneto-rotational instability that is believed to be the physical mechanism behind this transport yield values of α of roughly 0.1–0.2, consistent with values determined from observations of accreting white dwarfs⁴. Equivalent viscosity parameters have hitherto not been estimated for disks around neutron stars or black holes. Here we report the results of an analysis of archival X-ray light curves of 21 outbursts in black-hole X-ray binaries. By applying a Bayesian approach to a model of accretion, we determine corresponding values of α of around 0.2–1.0. These high values may be interpreted as an indication either of a very high intrinsic rate of angular-momentum transport in the disk, which could be sustained by the magneto-rotational instability only if a large-scale magnetic field threads the disk^{5–7}, or that mass is being lost from the disk through substantial outflows, which strongly shape the outburst in the black-hole X-ray binary. The lack of correlation between our estimates of α and the accretion state of the binaries implies that such outflows can remove a substantial fraction of the disk mass in all accretion states and therefore suggests that the outflows correspond to magnetically driven disk winds rather than thermally driven ones, which require specific radiative conditions⁸.

The disk-instability model^{9–14} was developed to explain outbursts in compact binaries in which a white dwarf accretes from a low-mass companion¹⁵. A cool (neutral) quiescent disk is built up through steady mass transfer from the companion star, causing the temperature of the disk to increase. At some radius (called the ignition radius), the temperature of the disk will eventually reach the temperature at which hydrogen ionizes. This triggers a thermal–viscous instability within the disk, due to the steep temperature dependence of opacity in this temperature range. As a result, the disk cycles between a hot, ionized, outburst state and a cold, neutral, quiescent state. The growth of the thermal–viscous instability at the ignition radius results in two heating fronts propagating inwards and outwards through the disk. This propagation brings the disk into a hot state, causing rapid in-fall of matter onto the compact object and a bright optical and ultraviolet outburst.

As the disk is depleted over time (because mass falls onto the compact stellar remnant at a higher rate than it is being transferred from the companion star), the temperature and mass-accretion rate in the outer radii will eventually be reduced to the point at which hydrogen can recombine. This triggers the formation and propagation of a cooling front that returns the disk to its quiescent (neutral) state. This predicted behaviour, which is characterized by alternating periods of outbursts and quiescence, matches observations of accreting white dwarfs well.

However, changes to the theory are needed for the close binaries known as low-mass X-ray binaries, which contain more compact stellar remnants (such as neutron stars and stellar-mass black holes).

There are 18 confirmed black-hole low-mass X-ray binaries in our Galaxy, identified through bright X-ray outbursts which indicate rapid accretion episodes¹⁶. These outbursts¹⁶ last considerably longer, and recur much less frequently, than those from many types of accreting white dwarf, owing to heating of the outer disk by X-rays emitted in the inner regions of the accretion flow¹⁷.

X-ray irradiation keeps the accretion disk in its hot (ionized) state over the viscous timescale. This timescale, which is encoded in observed outburst light curves, is related to the efficiency of angular-momentum transport directly and thus provides a means of measuring this efficiency. See Methods and Extended Data Fig. 1 for a detailed discussion of the characteristic three-stage outburst decay profile present in the light curve of a black-hole low-mass X-ray binary.

The magneto-rotational instability is thought to provide the physical mechanism behind angular-momentum (and mass) transport in accretion disks¹⁸. The effective viscosity in these disks, which is commonly parameterized using the α viscosity prescription¹⁹, encapsulates the efficiency of this transport process. Physically, the viscosity parameter α sets the viscous time of the accretion flow through the disk and thus, according to the disk-instability model, is encoded within the decay profile of the light curve of an outburst. A disk with higher viscosity (higher α) in outburst will accrete mass more quickly, resulting in shorter decay times and shorter outburst durations²⁰.

The α viscosity has been inferred only in (non-irradiated) disks around accreting white dwarfs, by comparing the outburst timescales from observed light curves to synthetic model light curves created by numerical disk codes for different α inputs⁴. Values of α have not previously been measured in irradiated accretion disks, such as those around stellar-mass black holes in low-mass X-ray binaries. The assertion²¹ that $\alpha \approx 0.2$ –0.4 in such systems was deduced from calculations²⁰ of “detailed models of complete light curves”, not from detailed comparison of models with observations. Note that we learned of a recent study²² of the black-hole low-mass X-ray binary 4U 1543–475 after acceptance of this manuscript.

Accordingly, we have built a Bayesian approach that characterizes the angular-momentum (and mass) transport that occurs in disks in low-mass X-ray binary systems. The α viscosity parameter in a hot, outbursting disk (α_h) sets the timescale on which matter moves through the hot (ionized) portion of the disk and thus controls the duration of the first stage of the decay profile observed in an X-ray light curve (see Methods for details). This (viscous) timescale varies depending on the mass of the compact object and the size of the accretion disk, with the size of the disk itself governed by the ratio of component masses in the system and the orbital period of the binary. To reconcile the multi-level, interconnected relationships that exist between these parameters that define the properties of the accretion flow, we use Bayesian

¹Department of Physics, University of Alberta, CCIS 4-181, Edmonton, Alberta T6G 2E1, Canada. ²Institut d'Astrophysique de Paris, CNRS et Sorbonne Universités, UPMC Paris 06, UMR 7095, 98bis Boulevard Arago, 75014 Paris, France. ³Nicolaus Copernicus Astronomical Centre, Polish Academy of Sciences, ulica Bartycka 18, 00-716 Warsaw, Poland. ⁴Université Grenoble Alpes, CNRS, Institut de Planétologie et d'Astrophysique de Grenoble (IPAG), F-38000 Grenoble, France.

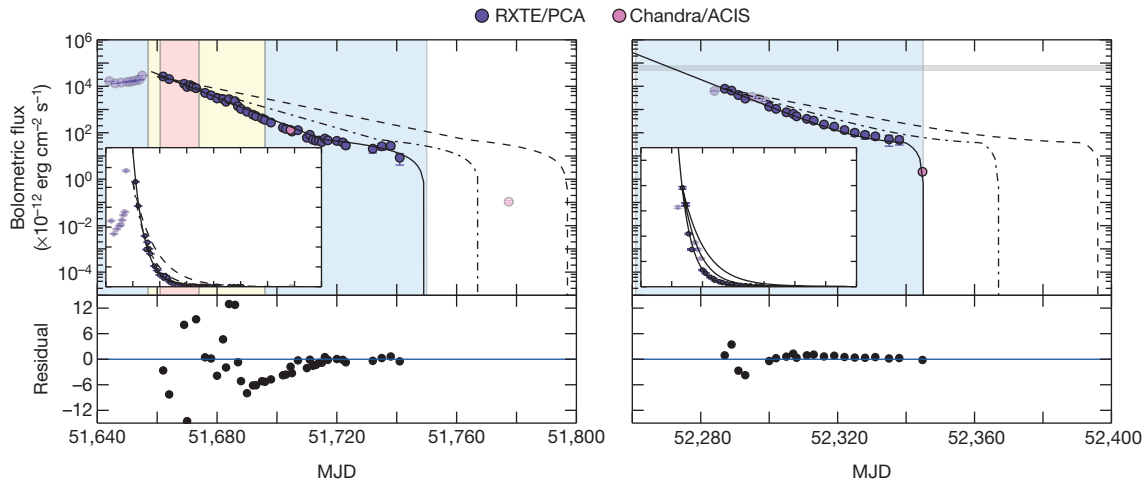


Figure 1 | Example light curves of outbursts in low-mass X-ray binaries.

The figure displays the observed bolometric X-ray light curves for the 2000 (left) and 2001–2002 (right) outbursts of the low-mass X-ray binary XTE J1550–564, which harbours a black hole with a mass of 10.4 ± 2.3 solar masses (M_{\odot})¹⁶. In this source, which has undergone multiple outbursts in the past two decades, we measure an extremely high value of the viscosity parameter α . Shading in the background indicates the accretion state of the source during the outbursts: blue, hard; yellow, intermediate; red, soft. Although XTE J1550–564 transitions from the soft to hard accretion state during the decay of the 2000 outburst, the light curve shows no signature of that transition. Disk outflows have been observed only in the soft and intermediate states, or at high flux levels (greater than 10% Eddington)²⁷, above the grey bar (left) in the hard state.

hierarchical modelling (see Methods of details). This technique allows us to derive: (i) the timescales associated with individual stages of the decay of the outburst, and (ii) the rate of mass accretion through the disk during, and the time of occurrence of, the transitions between the individual decay stages. Ultimately, the Bayesian technique allows us to take into account effectively our prior knowledge of the orbital parameters that define a low-mass X-ray binary system (black-hole mass, companion mass and orbital period) and thus to sample α_h directly from its observed X-ray light curve.

We analysed a representative sample of X-ray light curves of 21 individual outbursts of 12 black-hole low-mass X-ray binary systems from the WATCHDOG project¹⁶ (Extended Data Table 1). In Fig. 1 we show examples of the analytical irradiated-disk-instability model fitted to observed data. In this figure, we overlay predicted decay profiles that illustrate the way in which varying α_h changes the predicted light-curve decay profile. For these 21 outbursts, we derive $0.19 < \alpha_h < 0.99$ (see Fig. 2 and Extended Data Table 2). These results represent a derivation of α in accretion disks of low-mass X-ray binaries from a fit to the observed light curves of outbursts in such systems.

There are two possible explanations for the high values of α that we measure. The first is that we really are measuring the intrinsic value of α for these disks. The only way to reproduce such high intrinsic values of α in accretion-disk simulations is for a net magnetic field to thread the disk, with concurrent mass outflows strongly shaping the outburst as a whole. Simulations of angular-momentum transport driven by the magneto-rotational instability, carried out in vertically stratified boxes that represent a local patch of the disk (shearing box), typically yield $\alpha \approx 0.02$ without a net magnetic flux^{23,24}. Convection enhances transport to yield $\alpha \approx 0.2$ in the conditions appropriate to accreting white dwarfs^{1–3}. This value is consistent with those deduced from observations of outbursts in the non-irradiated disks around these objects⁴, but is insufficient to explain the higher values of $\alpha \gtrsim 0.2$ that we measure from outbursts in black holes. However, when the shearing box is threaded by a net magnetic flux, simulations show that α scales as $\beta^{-1/2}$ (where β is the ratio of the thermal pressure

Coloured circles represent data from different X-ray instruments: the Proportional Counter Array aboard the Rossi X-ray Timing Explorer (RXTE/PCA), or the Advanced CCD Imaging Spectrometer aboard the Chandra X-ray Observatory (Chandra/ACIS). Translucent data indicate the rise of the outburst, which was not included in the fits. Error bars show the statistical uncertainties of the instruments. The insets show the outbursts on a linear scale. The best-fitting analytical model (solid black line) and residuals (lower panel) are displayed in both panels. We measure $\alpha_h = 0.96 \pm 0.15$ and $\alpha_h = 0.99^{+0.15}_{-0.14}$ from the light curves of the 2000 and 2001–2002 outbursts, respectively. We over-plot the resulting decay profiles corresponding to $\alpha_h = 0.7$ (dot-dashed line) and $\alpha_h = 0.5$ (dashed line), demonstrating the way in which the shape of the light curve changes with different values of α . MJD, modified Julian date.

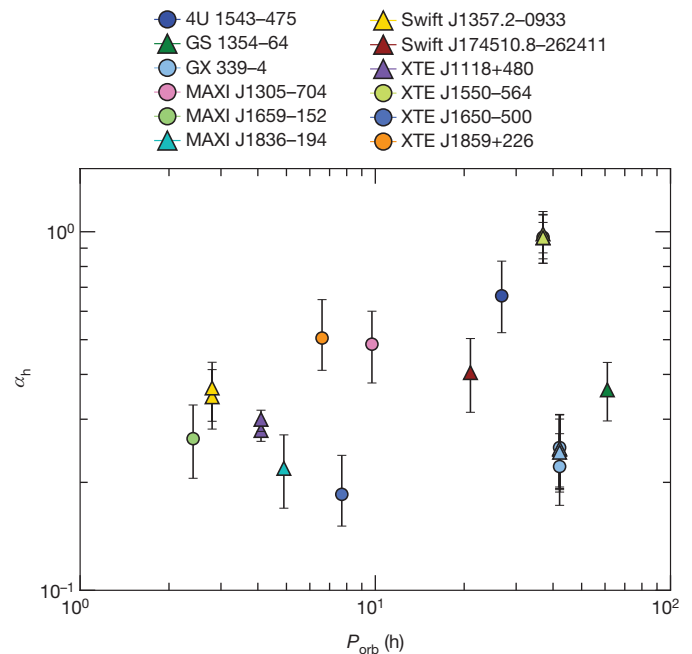


Figure 2 | Characterization of the mass-transport process in accretion disks. The α viscosity parameter in the hot, ionized zone of the disk (α_h), which encompasses the efficiency of angular-momentum (and mass) transport in accretion disks, as derived using our Bayesian methodology, is plotted against the orbital period (P_{orb}) of the binary system, for the 21 individual outbursts that occurred in our sample of 12 Galactic black-hole low-mass X-ray binaries with measured orbital periods. The different colours represent individual sources. Error bars represent the 68% confidence interval. The values of α_h are derived both for outbursts during which the source cycles through all accretion states (hard, intermediate and soft; circles) and for those during which the source remains in the hard accretion state (triangles).

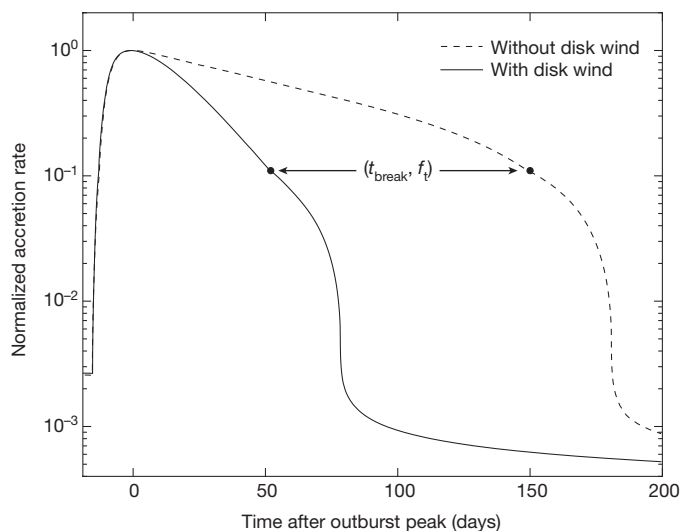


Figure 3 | Toy model of a disk ‘wind’. Two model light curves for an irradiated disk with $\alpha_h = 0.2$ around a $6M_\odot$ black hole are shown: dashed line, assuming no mass loss; solid line, including a term to account for mass loss during the outburst. The latter is computed assuming mass loss is proportional to the central mass-accretion rate onto the black hole ($\dot{M}_w = \varepsilon_w \dot{M}_c$) during the decay (meant to be representative of a disk-wind-type outflow). Although the shape of the profile remains the same, the effective timescale τ_e is reduced to $(1 - \varepsilon_w)\tau_e$. Thus, as the fraction of mass lost increases, τ_e decreases, mimicking the effect an arbitrary large value of α has on the light-curve profile (that is, high α corresponds to fast decay). A measurement of $\alpha = 1$ would correspond to a disk with $\alpha = 0.2$ and $\varepsilon_w = 0.8$ in the toy model, indicative of a substantial outflow. Note that, although this model assumes that the local outflow rate is related to the local accretion rate in the disk, this need not be the case. Further, this simplifying assumption, used purely to solve for the light curve numerically, limits what we can say about how much mass is lost in the outflow. This model requires $\dot{M}_w/\dot{M}_c < 1$, but it is certainly possible that the outflow rate is larger than the central mass-accretion rate onto the black hole ($\dot{M}_w/\dot{M}_c > 1$). The transition (which occurs at a flux f_l and time t_{break}) between the viscous and irradiation-controlled stages of the decay in each light-curve is indicated by a black filled circle.

to the imposed magnetic pressure), reaching values as high as $\alpha \approx 1$ when β is roughly $1-10^3$, with $\beta = 1$ the lower limit for the magnetorotational instability to operate in a thin disk⁵⁻⁷. Hence, strong intrinsic angular-momentum transport indicates the presence of a large-scale field in the accretion disk, origin and evolution of which have yet to be determined in black-hole transients. Moreover, simulations that reproduce high intrinsic α also display strong outflows, which actually do not remove much angular momentum; thus, angular-momentum transport is still primarily driven by the magneto-rotational instability.

The second possibility is that the intrinsic values of α for accretion disks of black-hole low-mass X-ray binaries are smaller than we measure (for example, comparable to 0.2) and that unspecified strong mass outflows are shaping the overall light-curve profiles that we observed. Figure 3 illustrates how including a term to account for mass (and angular-momentum) loss within the irradiated-disk-instability model results in a model that mimics the effect that high α has on the light-curve decay profile.

In both cases, substantial outflows appear to have a key role in regulating the disk-accretion process. Strong mass outflows have been observed in outbursting low-mass X-ray binaries in the soft or intermediate accretion states, or at high flux levels (greater than 10% Eddington) in the hard accretion state²⁵⁻²⁷, in the form of accretion-disk winds. These outflows have been attributed to thermal winds driven by X-ray irradiation or to magnetic winds driven by centrifugal acceleration along magnetic field lines anchored in the disk^{8,28}. It has recently been shown that thermally driven winds (such as Compton-heated

winds²⁶) can be produced only in the soft accretion state, because the ionization state of the wind becomes unstable in the hard accretion state (see, for example, refs 29, 30). The absence of a correlation between the values of α and the X-ray flux or accretion state in our outburst sample suggests that the outflow mechanism is generic, and favours magnetically driven over thermally driven outflows.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 September; accepted 17 November 2017.

Published online 22 January 2018.

- Hirose, S., Blaes, O., Krolik, J., Coleman, M. & Sano, T. Convection causes enhanced magnetic turbulence in accretion disks in outburst. *Astrophys. J.* **787**, 1 (2014).
- Coleman, M., Kotko, I., Blaes, O., Lasota, J.-P. & Hirose, S. Dwarf nova outbursts with magnetorotational turbulence. *Mon. Not. R. Astron. Soc.* **462**, 3710–3726 (2016).
- Scepi, N., Lesur, G., Dubus, G. & Flock, M. Impact of convection and resistivity on angular momentum transport in dwarf novae. *Astron. Astrophys.* <https://doi.org/10.1051/0004-6361/201731900> (2017).
- Kotko, I. & Lasota, J.-P. The viscosity parameter and the properties of accretion disc outbursts in close binaries. *Astron. Astrophys.* **545**, A115 (2012).
- Lesur, G., Ferreira, J. & Ogilvie, G. I. The magnetorotational instability as a jet launching mechanism. *Astron. Astrophys.* **550**, A61 (2013).
- Bai, X.-N. & Stone, J. Local study of accretion disks with a strong vertical magnetic field: magnetorotational instability and disk outflow. *Astrophys. J.* **767**, 30–48 (2013).
- Salvesen, G., Simon, J. B., Armitage, P. J. & Begelman, M. C. Accretion disc dynamo activity in local simulations spanning weak-to-strong net vertical magnetic flux regimes. *Mon. Not. R. Astron. Soc.* **457**, 857–874 (2016).
- Higginbottom, N. & Proga, D. Coronae and winds from irradiated disks in x-ray binaries. *Astrophys. J.* **807**, 107–116 (2015).
- Osaki, Y. An accretion model for the outbursts of U Geminorum stars. *Publ. Astron. Soc. Jpn* **26**, 429–436 (1974).
- Meyer, F. & Meyer-Hofmeister, E. On the elusive cause of cataclysmic variable outbursts. *Astron. Astrophys.* **104**, 10–12 (1981).
- Smak, J. Accretion in cataclysmic binaries. IV. Accretion disks in dwarf novae. *Acta Astron.* **34**, 161–189 (1984).
- Faulkner, J., Lin, D. N. C. & Papaloizou, J. On the evolution of accretion disc flow in cataclysmic variables – I. The prospect of a limit cycle in dwarf nova systems. *Mon. Not. R. Astron. Soc.* **205**, 359–375 (1983).
- Huang, M. & Wheeler, J. Thermal instability accretion disc model for the X-ray transient A0620–00. *Astrophys. J.* **343**, 229–240 (1989).
- Cannizzo, J. K. *Accretion Disks in Compact Stellar Systems* (World Scientific, 1993).
- Warner, B. *Cataclysmic Variable Stars* Ch. 3, 126–215 (Cambridge Univ. Press, 1995).
- Tetarenko, B., Sivakoff, G., Heinke, C. & Gladstone, J. C. Watchdog: a comprehensive all-sky database of galactic black hole x-ray binaries. *Astrophys. J. Suppl. Ser.* **222**, 15 (2016).
- van Paradijs, J. On the accretion instability in soft x-ray transients. *Astrophys. J.* **464**, L139–L141 (1996).
- Balbus, S. & Hawley, J. Instability, turbulence, and enhanced transport in accretion disks. *Rev. Mod. Phys.* **70**, 1–53 (1998).
- Shakura, N. I. & Sunyaev, R. A. Black holes in binary systems. Observational appearance. *Astron. Astrophys.* **24**, 337–355 (1973).
- Dubus, G., Hameury, J.-M. & Lasota, J.-P. The disc instability model for x-ray transients: evidence for truncation and irradiation. *Astron. Astrophys.* **373**, 251–271 (2001).
- King, A. R., Pringle, J. E. & Livio, M. Accretion disc viscosity: how big is alpha? *Mon. Not. R. Astron. Soc.* **376**, 1740–1746 (2007).
- Lipunova, G. V. & Malanchev, K. L. Determination of the turbulent parameter in accretion discs: effects of self-irradiation in 4U 1543–47 during the 2002 outburst. *Mon. Not. R. Astron. Soc.* **468**, 4735–4747 (2017).
- Davis, S. W., Stone, J. M. & Pessah, M. E. Sustained magnetorotational turbulence in local simulations of stratified disks with zero net magnetic flux. *Astrophys. J.* **713**, 52–65 (2010).
- Simon, J. B., Beckwith, K. & Armitage, P. J. Emergent mesoscale phenomena in magnetized accretion disc turbulence. *Mon. Not. R. Astron. Soc.* **422**, 2685–2700 (2012).
- Miller, J. M. et al. Simultaneous Chandra and RXTE spectroscopy of the microquasar H1743–322: clues to disk wind and jet formation from a variable ionized outflow. *Astrophys. J.* **646**, 394–406 (2006).
- Ponti, G. et al. Ubiquitous equatorial accretion disc winds in black hole soft states. *Mon. Not. R. Astron. Soc.* **422**, L11–L15 (2012).
- Neilsen, J. The case for massive, evolving winds in black hole x-ray binaries. *Adv. Space Res.* **52**, 732–739 (2013).
- Ohsuga, K. & Mineshige, S. Global structure of three distinct accretion flows and outflows around black holes from two-dimensional radiation-magnetohydrodynamic simulations. *Astrophys. J.* **736**, 2 (2011).

29. Chakravorty, S., Lee, J. C. & Neilsen, J. The effects of thermodynamic stability on wind properties in different low-mass black hole binary states. *Mon. Not. R. Astron. Soc.* **436**, 560–569 (2013).
30. Bianchi, S., Ponti, G., Muñoz-Darias, T. & Petrucci, P.-O. Photoionization instability of the Fe K absorbing plasma in the neutron star transient AX J1745.6–2901. *Mon. Not. R. Astron. Soc.* **472**, 2454–2461 (2017).

Acknowledgements B.E.T. thanks participants of the ‘disks17: Confronting MHD Theories of Accretion Disks with Observations’ programme, held at the Kavli Institute for Theoretical Physics (KITP), for their feedback and comments on this project, especially A. Veledina for advice regarding the analysis of the X-ray light curves and P. Charles for comments on the manuscript. B.E.T., G.R.S. and C.O.H. acknowledge support by NSERC Discovery Grants, and C.O.H. by a Discovery Accelerator Supplement. This research was supported in part by the National Science Foundation under grant number NSF PHY-1125915, via support for KITP. J.-P.L. acknowledges support by the Polish National Science Centre OPUS grant 2015/19/B/ST9/01099. J.-P.L. and G.D. also acknowledge support from the French Space Agency CNES. This research has made use of data, software, and/or web tools obtained from the High Energy Astrophysics Science Archive Research Center (HEASARC), a service of the Astrophysics Science Division at NASA/GSFC and of the Smithsonian Astrophysical Observatory’s High Energy Astrophysics Division, data supplied by the UK Swift Science Data Centre at the University of Leicester, and data provided by RIKEN,

JAXA and the MAXI team. This work has also made extensive use of NASA’s Astrophysics Data System (ADS).

Author Contributions B.E.T. performed the analysis of the X-ray data, wrote the Markov chain Monte Carlo light-curve-fitting algorithm and performed the light-curve fitting, built the Bayesian hierarchical methodology and wrote the paper. J.-P.L. helped to formulate the analytical version of the irradiated-disk-instability model that was fitted to the X-ray light curves, contributed to the interpretation of the data and assisted in writing the discussion in the paper. C.O.H. assisted in the analysis of the X-ray data and the light-curve-fitting process, and contributed to the interpretation of the data. G.D. contributed to the interpretation of the data and assisted in writing the discussion in the paper. G.R.S. assisted in writing the paper and contributed to the interpretation of the data.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher’s note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to B.E.T. (btetaren@ualberta.ca).

Reviewer Information *Nature* thanks D. Proga and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Archival X-ray data collection and reduction. We have collected all outburst data available since 1996, for each of the 12 systems in our source sample, from (i) the Proportional Counter Array (PCA) aboard the Rossi X-ray Timing Explorer (RXTE), (ii) the X-ray Telescope (XRT) aboard the Swift Observatory, (iii) the Gas-Slit Camera (GSC) aboard the Monitor of All-sky Image (MAXI) Telescope, (iv) the Advanced CCD Imaging Spectrometer (ACIS-S) and High Resolution Camera (HRC-S) aboard the Chandra X-ray Observatory, and (v) the European Photon Imaging Camera (EPIC) aboard XMM-Newton.

All X-ray light-curve data from RXTE/PCA was collected from the WATCHDOG project¹⁶. These authors compiled all available good pointed PCA observations (no scans or slews) from the HEASARC archive, for 77 black-hole X-ray binary sources in the Galaxy, over the entire 16-year RXTE mission. For each individual source in our sample, we use scripts from the WATCHDOG project, including the *rex* script within the Heasoft Software Package (<http://heasarc.nasa.gov/lheasoft/>), to reduce and extract (mission-long) daily time-binned, background-subtracted light curves in the 2–10-keV band from the PCA Std2 data available on that source in the WATCHDOG database. We also compiled all available MAXI/GSC data using the WATCHDOG project's online light-curve tool (<http://astro.physics.ualberta.ca/WATCHDOG>). This tool compiles all of the publicly available data from the MAXI archive (<http://maxi.riken.jp/top/>) in three standard bands (2–4 keV, 4–10 keV, 10–20 keV) and runs it through the WATCHDOG processing pipeline¹⁶. Using this tool, we extracted (mission-long) daily time-binned, background-subtracted light curves in the 2–10-keV band for each individual source (where available).

In addition, we use the Swift/XRT online product builder^{31,32} (http://www.swift.ac.uk/user_objects/index.php) to compile (mission-long) daily time-binned, background-subtracted light curves in the 2–10-keV band, using all available windowed-timing- and photon-counting-mode XRT pointed observations. Last, we collected all available Chandra/ACIS-S, Chandra/HRC-S and XMM-Newton/EPIC pointed observations from the literature for individual outbursts (where available). We then convert individual instrument count rates to fluxes in the 2–10-keV band using PIMMS v4.8c (<http://cxc.harvard.edu/toolkit/pimms.jsp>) and the spectral information available in the literature.

Conversion from count rate to bolometric flux. We use crabs as a baseline unit of flux to calculate approximate count-rate equivalences in the 2–10-keV-band data from RXTE/PCA, Swift/XRT and MAXI/GSC. Integration of the now-accepted, 'canonical', simple power-law spectrum of the Crab Nebula³³ over the 2–10-keV band gives us a straightforward method for converting between count rate and flux in this band. Assuming that a source spectrum is Crab-like in nature results in uncertainty in the computed source flux. However, given that it has been found that assuming a Crab-like spectral shape in narrow X-ray energy bands (such as the 2–10-keV band we make use of here) produces no more than a 20% (and typically less than 10%) error in the source flux for a flat power law versus a blackbody¹⁶, this approach is justified.

To convert flux in the 2–10-keV band to bolometric flux, we use the following bolometric corrections (BCs), estimated for each individual accretion state³⁴ that occurs during outbursts of black-hole low-mass X-ray binaries: hard state, BC = 5; soft and intermediate states, BC = 1.25. By combining the bolometric corrections discussed above with the daily accretion-state information for each outburst, obtained from the WATCHDOG project's¹⁶ online Accretion-State-By-Day tool (<http://astro.physics.ualberta.ca/WATCHDOG>), we are able to compute daily time-binned bolometric light curves.

Markov chain Monte Carlo (MCMC) fitting algorithm. We use a Bayesian approach to estimate the five parameters that describe the shape of an observed light-curve decay profile: (i) the exponential (viscous) decay timescale (τ_e), (ii) the linear (irradiation-controlled) decay timescale (τ_l), (iii) the X-ray flux of the system at the transition between exponential and linear decay stages (f_t), (iv) the time after the outburst peak when the transition between exponential and linear decay stages occurs (t_{break}), and (v) the X-ray flux limit of the exponential decay stage (f_2); see Extended Data Fig. 1. Using the *emcee* python package³⁵, an implementation of the Affine Invariant MCMC Ensemble Sampler³⁶, we apply a MCMC algorithm to fit the exponential (viscous) and linear (irradiation-controlled) stages of each decay simultaneously (as described in the main text and Extended Data Fig. 1), where applicable.

Before fitting occurs, secondary maxima and other rebrightening events^{20,37,38} that contaminate the decays are removed by hand. These data are not included in the fits; analysis of these rebrightening events will be presented in a later paper. The removal of these rebrightening events has no effect on the determination of α from the X-ray light curves. The remaining data are then fitted in logarithmic (bolometric) flux space with our five-parameter analytical model (see below for details).

The *emcee* python package runs a modified version of the Metropolis–Hastings algorithm, in which an ensemble of 'walkers' move simultaneously through and explore the parameter space. To fit each light curve, we use 50 walkers—10 times the dimension of model. For *emcee* to run optimally, we first set the initial positions of our ensemble of walkers appropriately in parameter space. To do so, we use *pyHarmonySearch*³⁹, an implementation of the harmony search global optimization algorithm, to perform an initial survey of our parameter space. *pyHarmonySearch* acts essentially as a less-time-consuming version of a brute-force grid-search method, allowing us to place our ensemble of walkers in a tight ball around the best guess that it finds. This best guess provides a starting point for the MCMC algorithm.

Prior distributions for each of the five parameters are also set from the results of *pyHarmonySearch*. In the case of a well-sampled light curve (near-continuous daily data throughout the outburst), a Gaussian prior for each parameter with a mean set by the results of *pyHarmonySearch* is used. In the case in which only scattered data are available on only a portion of the full decay, wide flat priors (based on expectations from other outbursts of the same source, or outbursts from sources with similar orbital periods) are used for each parameter.

After initialization, we begin running the MCMC algorithm on each light curve with a 500-step burn-in phase. Here the ensemble of walkers are evolved over a series of steps to ensure that the initial configuration that we have set enables the walkers to explore the parameter space sufficiently. At the end of the burn-in phase, if the initial configuration is appropriate for the problem, the walkers will have ended up in a high-probability region, a place in parameter space in which the states of the Markov chain are more representative of the distribution being sampled. After this phase, the MCMC algorithm is restarted, with the walkers starting at the final position they acquired during the burn-in phase, and run until convergence. The number of steps required for convergence depends on the amount of data available and the complexity of the decay profile of the outburst.

After likelihood maximization is performed, the MCMC algorithm outputs the converged solution in the form of posterior distributions of each parameter. We take the best-fit result (the best-fit value along with the upper and lower limits on this value) as the median and 1σ (68%) confidence interval of each posterior distribution, respectively.

The analytical outburst decay model. Extended Data Fig. 1 shows the predicted characteristic three-stage decay profile shape present in the light curve of a black-hole low-mass X-ray binary^{20,37,40}.

In the first stage (viscous decay), X-ray irradiation keeps the whole disk in a hot (ionized) state, preventing the formation of a cooling front. As more mass is accreted onto the black hole than is transferred from the companion at this time, the disk is drained by viscous accretion of matter only, resulting in an exponential-shaped decay profile on the viscous timescale. Eventually, as the mass in the disk and the mass transfer rate decrease, the dimming X-ray irradiation can no longer keep the outer regions of the disk in the hot (ionized) state and a cooling front forms, behind which the cold matter slows its inward flow drastically. At this point, the system enters the second stage (irradiation-controlled decay), during which the propagation of the cooling front is controlled by the decay of the irradiating X-ray flux. The hot (ionized) portion of the disk continues to flow and accrete, but gradually shrinks in size, causing a linear-shaped decay profile. Eventually, the mass accretion rate onto the black hole becomes small enough that X-ray irradiation no longer has a role. In this third and final stage (thermal decay), the cooling front propagates inward freely on a thermal–viscous timescale, resulting in a steeper linear decay in the light curve down to the quiescent accretion level.

The analytical model that we use to describe the outburst decay profiles in the light curves of black-hole low-mass X-ray binaries, predicted by the (irradiated) disk-instability model, is rooted in the 'classic' King and Ritter formalism³⁷. This formalism combines knowledge of the peak X-ray flux and of the outer radius of the irradiated disk to predict the shape that the decay of an X-ray light curve of a transient low-mass X-ray binary system would follow.

The temperature of most of the accretion disk in transient low-mass X-ray binaries during outburst is dominated by X-ray heating from the inner accretion region. The X-ray light curve will show an exponential decline if irradiation by the central X-ray source is able to ionize the entire disk, keeping it in the hot (ionized) state and preventing the formation of the cooling front⁴¹. The X-ray light curve will show a linear decline if irradiation by the central X-ray source is able to keep only a portion of the entire disk in the hot (ionized) state. In this case, the central X-ray flux will no longer be able to keep the outer regions of the disk above the hydrogen ionization temperature (around 10^4 K) and a cooling front will appear and propagate down the disk. Because the cooling front cannot move inward on a viscous timescale (the farthest it can move inward is set by the radius at which $T = 10^4$ K), a linear-shaped decline is observed in the light curve.

By assuming, as do many studies of X-ray irradiated disks in close binary systems, an isothermal disk model (that is, the disk is assumed to be vertically isothermal because it is irradiated, with the central mid-plane temperature equal to the effective temperature set by the X-ray irradiation flux at the disk surface⁴²), the critical X-ray luminosity for a disk radius R_{11} (in units of 10^{11} cm) has been derived³⁷ to be $L_{\text{crit,BH}} = 1.7 \times 10^{37} R_{11}^2 \text{ erg s}^{-1}$, above which the light curve should display an exponential decay shape and below which the light curve should display a linear decay shape. In this formalism, a well-sampled light curve (in both time and amplitude) should show a combination of exponential- and linear-shaped stages in the decay profile. The exponential decay is replaced with a linear decay when the X-ray flux has decreased sufficiently, resulting in a distinct brink (such as a break in slope) in the light-curve shape. By deriving analytical expressions for the shape that light-curve decays of transient low-mass X-ray binaries systems take, the timescales of the exponential and linear stages of a decay, the peak mass-accretion rate (and in-turn X-ray luminosity for a given accretion efficiency) and the time at which the exponential decay is replaced by the linear decay have been predicted³⁷.

This approach has since been validated by smooth-particle-hydrodynamics accretion simulations⁴³ and applied to observations of various classes of X-ray binaries^{44–49} with varied success. However, although the King and Ritter formalism has, coincidentally, been found to agree relatively well with observations, it oversimplifies the physics of the X-ray-irradiated disks to which it is applied⁴¹. Therefore, we instead use a modified version of the King and Ritter formalism.

In this modified version we (i) include the effects of continuing mass transfer from the donor star^{47,48}, and (ii) use a disk structure^{20,40} in which X-ray irradiation that affects the disks of black-hole low-mass X-ray binaries is modelled using a general irradiation law:

$$T_{\text{irr}}^4 = \frac{C_{\text{irr}} L_X}{4\pi\sigma R^2}$$

Here, the irradiation parameter C_{irr} is defined as the fraction of the central X-ray luminosity ($L_X = \eta c^2 \dot{M}_c$ for accretion efficiency η) that heats up the disk. Because C_{irr} contains information about the illumination and disk geometry, and the temperature profile of the X-ray irradiation, it effectively parameterizes our ignorance of how these disks are actually irradiated. Physically, C_{irr} controls the timescale of the linear decay stage (and the overall outburst duration) and when the transition between decay stages occur, and sets a limit on the amount of mass that can be accreted during the outburst. Stronger irradiation (larger C_{irr}) increases the duration of the outburst and thus the relative amount of matter able to be accreted during an outburst. Consequently, if more matter is accreted during outburst, the subsequent time in quiescence will lengthen because the disk will require more time to build up again.

Following the procedure outlined in previous work^{47,48}, but instead using the general irradiation law defined above, yields the following analytical form for the flux of a black-hole low-mass X-ray binary as a function of time during the exponential (viscous) and linear (irradiation-controlled) stages of the decay:

$$f_X = \begin{cases} (f_1 - f_2) \exp\left(-\frac{t - t_{\text{break}}}{\tau_e}\right) + f_2 \\ f_1 \left(1 - \frac{t - t_{\text{break}}}{\tau_l}\right) \end{cases}$$

Here τ_e and τ_l are defined as the viscous (exponential) decay timescale in the hot (ionized) zone of the disk and the linear decay timescale, respectively; $f_2 = \eta c^2 (-\dot{M}_2)/(4\pi d^2)$ is the flux limit of the exponential decay, which depends on the mass-transfer rate from the companion $-\dot{M}_2$ and the distance to the source d ; t_{break} is defined as the time when the temperature of the outer edge of the disk is just sufficient to remain in a hot (ionized) state; and f_1 is the corresponding X-ray flux of the system at time t_{break} . We perform fits in flux, as opposed to luminosity, space to avoid the correlated errors (due to an uncertain distance) that would arise if we were to fit the latter; the uncertain distance (and other parameters) are incorporated below.

By fitting this model to our sample of observed X-ray light curves we can derive the viscous decay timescales in black-hole low-mass X-ray binaries to range between roughly 50 and 190 days, consistent with previous conclusions⁵⁰; see Extended Data Table 2 for fit results.

The Bayesian hierarchical methodology. We quantify the angular-momentum (and mass) transport that occurs in the irradiated accretion disks in low-mass X-ray binary systems using the α viscosity parameter. In the current form of the disk-instability model, the use of this simple parameter results from the inability of current numerical simulations to follow *ab initio* turbulent transport driven by

the magneto-rotational instability on viscous timescales in a global model of the accretion disk.

This parameter is encoded within the viscous (exponential) stage of the light-curve decay profile. During this first stage of the decay, irradiation of the disk traps it in a hot (ionized) state that allows a decay of the central mass-accretion rate only on a viscous timescale: $\tau_e = R_{\text{h,disk}}^2/(3\nu_{\text{KR}})$, where ν_{KR} is the Shakura–Sunyaev viscosity¹⁹, the average value of the kinematic viscosity coefficient near the outer edge of the disk³⁷, and $R_{\text{h,disk}}$ is the radius of the hot (ionized) zone of the disk. For Keplerian disks, the Shakura–Sunyaev viscosity is related to the dimensionless viscosity parameter in the hot disk α_h by $\nu_{\text{KR}} = \alpha_h c_s^2/\Omega_k$, where Ω_k is the Keplerian angular velocity and c_s is the sound speed in a disk (proportional to $T_c^{0.5}$, where T_c is the central mid-plane temperature of the disk). Therefore, using $\Omega_k = (GM_1/R^3)^{1/2}$, the viscous timescale in the disk can be written as a function of α_h , the mass of the compact object M_1 and the radius of the accretion disk $R_{\text{h,disk}}$:

$$\left(\frac{\tau_e}{s}\right) = \left(\frac{G^{0.5} m_H M_\odot^{0.5} (10^6)}{3\gamma k_B T_c}\right) \left(\frac{\alpha_h}{0.1}\right)^{-1} \left(\frac{M_1}{M_\odot}\right)^{0.5} \left(\frac{R_{\text{h,disk}}}{10^{10} \text{ cm}}\right)^{0.5}$$

where G is the gravitational constant, m_h is the mass of a hydrogen atom, γ is the adiabatic index (the ratio of the specific heats of a gas at a constant pressure and a gas at a constant volume) and k_B is the Boltzmann constant. Because T_c is only weakly dependent on viscosity and X-ray irradiation in irradiated disks, we can approximate its value as a constant, 16,300 K (ref. 51).

Solving for α_h yields

$$\left(\frac{\alpha_h}{0.1}\right) = \left(\frac{G^{0.5} m_H M_\odot^{0.5} (10^6)}{3\gamma k_B T_c}\right) \left(\frac{\tau_e}{s}\right)^{-1} \left(\frac{M_1}{M_\odot}\right)^{0.5} \left(\frac{R_{\text{h,disk}}}{10^{10} \text{ cm}}\right)^{0.5}$$

Because α_h depends on parameters that characterize the outburst decay profile of a low-mass X-ray binary (that is, observed data) as well as on the orbital parameters that define the binary system (that is, parameters that we have prior knowledge of, namely, M_1 and $R_{\text{h,disk}}$, which is itself dependent on M_1 , the mass of the companion star in the system and the orbital period P_{orb}), we require a multi-level Bayesian statistical sampling technique to effectively sample α_h .

We therefore built a Bayesian hierarchical model. A Bayesian hierarchical model is a multi-level statistical model that enables the posterior distribution of some quantity to be estimated by integrating a combination of known prior distributions with observed data. In our case, the established orbital parameters of the binary (M_1 , binary mass ratio q , P_{orb}) for a system act as the known priors, and the quantitative outburst decay properties derived from fitting the light curves of a low-mass X-ray binary system with the analytical version of the irradiated disk-instability model that we developed (τ_e) act as the observed data.

Using *emcee*³⁵ (see above for details), our hierarchical model samples α_h simultaneously for all outbursts of each of the 12 sources in our sample using 240 walkers, 10 times the dimension of our model: 12 of these dimensions correspond to 6 established measurements of black-hole mass, 4 known binary mass ratios and 2 observationally based Galactic statistical population distributions (the Özel black-hole mass distribution⁵² and the distribution of binary mass ratios for the dynamically confirmed stellar-mass black holes in the Galaxy¹⁶); the remaining 12 dimensions correspond to the accretion disk radii for each system.

Initialization is accomplished by placing our ensemble of walkers in a tight ball around a best guess that corresponds to the best known estimates of the binary parameters (M_1 , q , P_{orb}) for each system. If a reliable estimate of M_1 is not known for a system, then the mean of the Ozel mass distribution⁵² is used. Similarly, if q is not known for a system, then the median of the uniform distribution between the minimum and maximum of the known values of mass ratio for all dynamically confirmed black holes in the Galaxy¹⁶ is used.

Our hierarchical model samples accretion disk radii from a uniform distribution between the circularization radius (R_{circ}) and the radius of the Roche lobe of the compact object (R_1) in the system, both of which depend on only M_1 , q and P_{orb} . Initial values of $R_{\text{h,disk}}$ are set as the median of the uniform distribution between R_{circ} and R_1 for each system, calculated using the best guess for M_1 and q (discussed above), and the known P_{orb} .

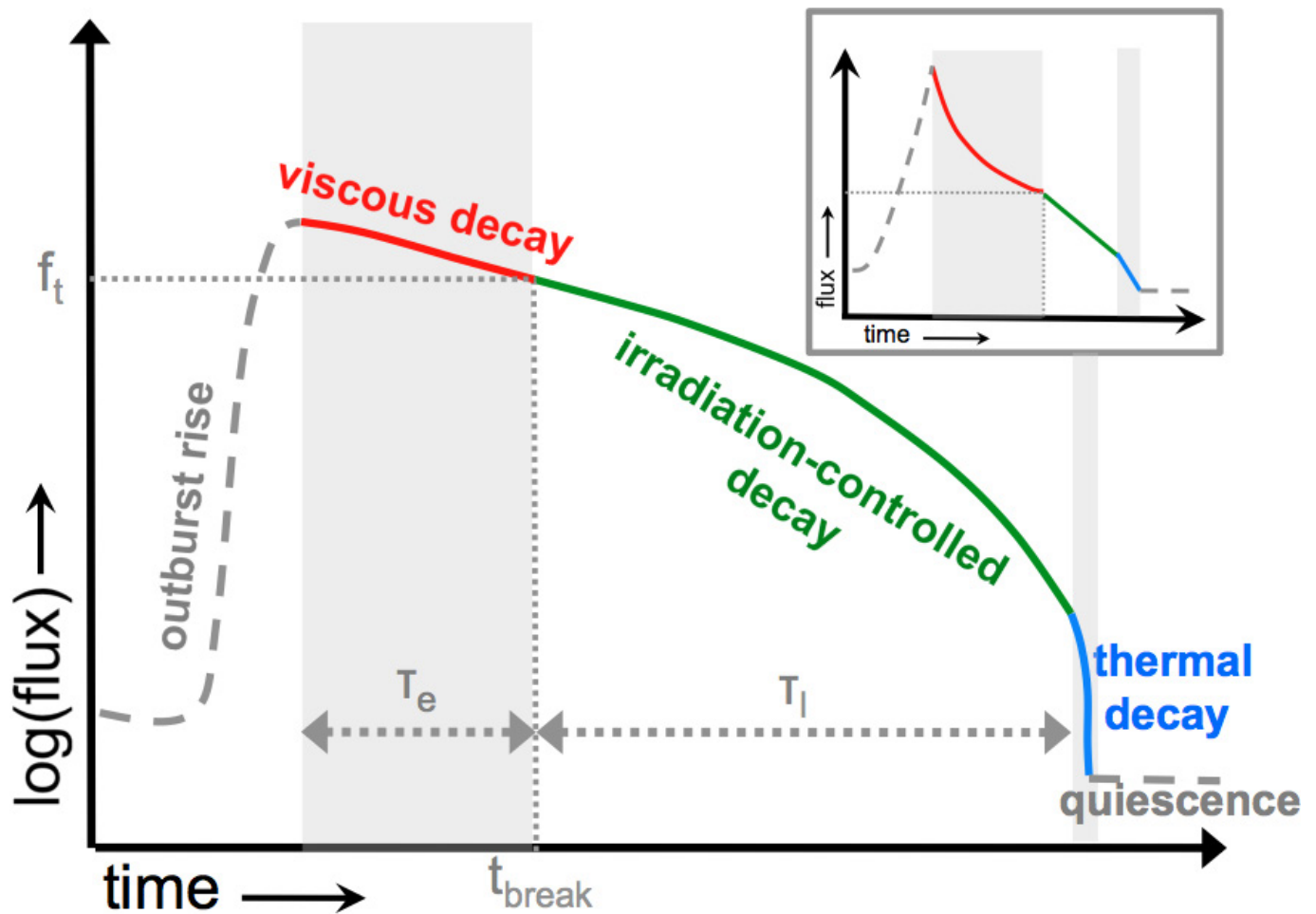
The prior distributions for each of the 24 parameters are also set using the best-guess orbital parameters for each binary system. If a constrained measurement of the parameter exists (that is, a value with uncertainty), then a Gaussian prior based on this measurement and its uncertainty is used. If only a range is quoted in the literature for a parameter, then a uniform prior is used. The prior distributions for $R_{\text{h,disk}}$ are taken as the uniform distribution between R_{circ} and R_1 for each system.

After initialization, we begin running *emcee* on the observed data (τ_e) with a 500-step burn-in phase. After this phase, *emcee* is restarted, with the walkers

starting at the final position they ended at in the burn-in phase, and run until convergence. Ultimately, *emcee* outputs the converged solution in the form of posterior distributions of α_{h} for each outburst or system. The converged value and upper and lower limits on this value are taken as the median and 1σ confidence interval of each posterior distribution, respectively; see Extended Data Table 2.

Data availability. The datasets generated during and analysed during this study are available from the corresponding author on reasonable request.

31. Evans, P. A. *et al.* Methods and results of an automatic analysis of a complete sample of Swift-XRT observations of grbs. *Mon. Not. R. Astron. Soc.* **397**, 1177–1201 (2009).
32. Evans, P. A. *et al.* An online repository of Swift/XRT light curves of gamma-ray bursts. *Astron. Astrophys.* **469**, 379–385 (2007).
33. Toor, A. & Seward, F. D. The Crab nebula as a calibration source for x-ray astronomy. *Astron. J.* **79**, 995–999 (1974).
34. Migliari, S. & Fender, R. Jets in neutron star x-ray binaries: a comparison with black holes. *Mon. Not. R. Astron. Soc.* **366**, 79–91 (2006).
35. Foreman-Mackey, D., Hogg, D. W., Lang, D. & Goodman, J. *emcee*: the MCMC hammer. *Publ. Astron. Soc. Jpn* **125**, 306–312 (2013).
36. Goodman, J. & Weare, J. Ensemble samplers with affine invariance. *Comm. App. Math. Comp. Sci.* **5**, 65–80 (2010).
37. King, A. R. & Ritter, H. The light curves of soft x-ray transients. *Mon. Not. R. Astron. Soc.* **293**, L42–L48 (1998).
38. Menou, K., Hameury, J.-M., Lasota, J.-P. & Narayan, R. Disc instability models for X-ray transients: evidence for evaporation and low α -viscosity? *Mon. Not. R. Astron. Soc.* **314**, 498–510 (2000).
39. Geem, Z., Kim, J. & Loganathan, G. A new heuristic optimization algorithm: harmony search. *Simulation* **76**, 60–68 (2001).
40. Dubus, G., Lasota, J.-P., Hameury, J.-M. & Charles, P. X-ray irradiation in low-mass binary systems. *Mon. Not. R. Astron. Soc.* **303**, 139–147 (1999).
41. Lasota, J.-P. The disc instability model of dwarf novae and low-mass x-ray binary transients. *New Astron. Rev.* **45**, 449–508 (2001).
42. de Jong, J., van Paradijs, J. & Augusteijn, T. Reprocessing of x rays in low-mass x-ray binaries. *Astron. Astrophys.* **314**, 484–490 (1996).
43. Truss, M. R., Wynn, G. A., Murray, J. R. & King, A. R. The origin of the rebrightening in soft x-ray transient outbursts. *Mon. Not. R. Astron. Soc.* **337**, 1329–1339 (2002).
44. Torres, M. A. P. *et al.* Observations of the 599 Hz accreting x-ray pulsar IGR J00291+5934 during the 2004 outburst and in quiescence. *Astrophys. J.* **672**, 1079–1090 (2008).
45. Šimon, V., Bartolini, C., Piccioni, A. & Guarenieri, A. Interpretation of the 1998 outburst of the unique X-ray transient CI Camelopardalis (XTE j0421+560). *Mon. Not. R. Astron. Soc.* **369**, 355–359 (2006).
46. Campana, S., Coti Zelati, F. & D'Avanzo, P. Mining the Aql X-1 long-term X-ray light curve. *Mon. Not. R. Astron. Soc.* **432**, 1695–1700 (2013).
47. Heinke, C. O., Bahramian, A., Degenaar, N. & Wijnands, R. The nature of very faint x-ray binaries: hints from light curves. *Mon. Not. R. Astron. Soc.* **447**, 3034–3043 (2015).
48. Powell, C., Haswell, C. & Falanga, M. Mass transfer during low-mass x-ray transient decays. *Mon. Not. R. Astron. Soc.* **374**, 466–476 (2007).
49. Shahbaz, T., Charles, P. & King, A. Soft x-ray transient light curves as standard candles: exponential versus linear decays. *Mon. Not. R. Astron. Soc.* **301**, 382–388 (1998).
50. Yan, Z. & Yu, W. X-ray outbursts of low-mass x-ray binary transients observed in the RXTE era. *Astrophys. J.* **805**, 87 (2015).
51. Lasota, J.-P., King, A. R. & Dubus, G. X-ray transients: hyper- or hypo-luminous? *Astrophys. J.* **801**, L4 (2015).
52. Özel, F., Psaltis, D., Narayan, R. & McClintock, J. The black hole mass distribution in the galaxy. *Astrophys. J.* **725**, 1918–1927 (2010).



Extended Data Figure 1 | Schematic light curve for an outburst of a low-mass X-ray binary system. The profile shown corresponds to the light curve predicted by the (irradiated) disk-instability model for an outbursting irradiated disk. τ_e and τ_l represent the timescales of the exponential (viscous) and linear (irradiation-controlled) decay stages in

the light curve, respectively. The time and flux at which the transition between the viscous and exponential stages of the decay occurs (marking the point at which the temperature in the outer disk drops below the ionization temperature of hydrogen) are represented by t_{break} and f_t , respectively. The inset shows the same light-curve profile on a linear scale.

Extended Data Table 1 | Binary orbital parameters for our Galactic black-hole low-mass X-ray binary source sample

Source Name	distance	M_1	q	P_{orb}
	(kpc)	(M_{\odot})	(M_2/M_1)	(hrs)
XTE J1118+480	1.72 ± 0.1	7.2 ± 0.72	0.024 ± 0.009	4.1
MAXI J1305–704	-	-	-	9.74
Swift J1357.2–0933	1.5–6.3	12.4 ± 3.6	-	2.8
GS 1354–64	-	-	0.12 ± 0.04	61.1
4U 1543–475	7.5 ± 0.5	9.4 ± 2.0	0.25–0.31	26.8
XTE J1550–564	4.4 ± 0.5	10.39 ± 2.3	0.031–0.037	37.0
XTE J1650–500	2.6 ± 0.7	4.7 ± 2.2	-	7.7
MAXI J1659–152	1.6–8.0	-	-	2.414
GX 339–4	8.0 ± 2.0	-	-	42.1
Swift J1745–26	-	-	-	≤ 21
MAXI J1836–194	-	-	-	< 4.9
XTE J1859+226	8 ± 3	10.83 ± 4.67	-	6.6

When no acceptable estimates of distance, black-hole mass M_1 or binary mass ratio q are available (indicated by dashes), known Galactic distributions^{16,52} are used.

Extended Data Table 2 | Quantities derived to describe the mass-transport process in the accretion disks of outbursting low-mass X-ray binaries

Source Name	Outburst ID	τ_e (days)	R_{disc} ($\times 10^{10}$ cm)	α_h	Accretion State(s) Reached	Max Eddington Fraction
4U1543–475	2002	58.94 ± 0.42	$27.91^{+12.2}_{-10.1}$	$0.66^{+0.16}_{-0.14}$	H,I,S	0.16
GS1354–64	2015	$139.69^{+0.63}_{-0.65}$	$52.10^{+18.1}_{-15.3}$	$0.362^{+0.070}_{-0.066}$	H	$0.085^{a,b}$
GX339–4	1996–1999	$167.17^{+2.12}_{-2.31}$	$37.30^{+15.6}_{-14.2}$	$0.250^{+0.059}_{-0.056}$	H,I,S	0.011^a
	2008	$168.24^{+5.89}_{-5.81}$	$37.24^{+15.5}_{-14.1}$	$0.247^{+0.061}_{-0.056}$	H	0.0059^a
	2009	$166.88^{+4.96}_{-4.48}$	$37.33^{+15.4}_{-14.1}$	$0.249^{+0.060}_{-0.057}$	H	0.0088^a
	2013	$172.37^{+3.14}_{-3.51}$	$37.29^{+15.2}_{-14.0}$	$0.242^{+0.058}_{-0.054}$	H	0.014^a
	2014/2015	$188.90^{+0.25}_{-0.23}$	$37.20^{+15.4}_{-14.0}$	$0.222^{+0.049}_{-0.052}$	H,I,S	0.15^a
MAXIJ1305–704	2012	$52.90^{+0.11}_{-0.12}$	$12.72^{+5.70}_{-4.65}$	$0.49^{+0.11}_{-0.11}$	H,I,S	$0.051^{a,b}$
MAXIJ1659–152	2010/2011	$60.69^{+1.19}_{-1.23}$	$5.494^{+2.31}_{-2.07}$	$0.265^{+0.059}_{-0.064}$	H,I,S	0.15^a
MAXIJ1836–194	2011/2012	$93.09^{+1.81}_{-2.00}$	$8.838^{+3.61}_{-3.34}$	$0.220^{+0.049}_{-0.053}$	H	$0.11^{a,b}$
SwiftJ1357.2–0933	2011	$68.31^{+2.16}_{-2.05}$	$6.850^{+3.00}_{-2.46}$	$0.346^{+0.067}_{-0.065}$	H	0.0019
	2017	$64.89^{+3.47}_{-3.68}$	$6.730^{+2.95}_{-2.42}$	$0.366^{+0.066}_{-0.070}$	H	0.00038
SwiftJ174510.8–262411	2012/2013	$81.49^{+1.92}_{-1.86}$	$23.49^{+9.71}_{-8.92}$	$0.410^{+0.097}_{-0.091}$	H	$1.2^{a,b}$
XTEJ1118+480	1999/2000	$85.96^{+0.55}_{-0.56}$	$12.94^{+1.80}_{-2.10}$	$0.279^{+0.017}_{-0.018}$	H	0.0017
	2005	$79.01^{+1.29}_{-1.04}$	$12.95^{+1.78}_{-2.10}$	$0.303^{+0.019}_{-0.021}$	H	0.00047
XTEJ1550–564	2000	$61.78^{+0.38}_{-0.37}$	$55.97^{+10.7}_{-9.53}$	$0.96^{+0.15}_{-0.16}$	H,I,S	0.043
	2001	$61.92^{+5.04}_{-5.79}$	$56.14^{+10.7}_{-9.68}$	$0.962^{+0.101}_{-0.089}$	H,I	0.0068
	2001/2002	$60.38^{+0.64}_{-0.63}$	$56.06^{+10.8}_{-9.60}$	$0.99^{+0.15}_{-0.15}$	H	0.013
	2003	$61.89^{+0.55}_{-0.52}$	$55.99^{+10.7}_{-9.52}$	$0.96^{+0.15}_{-0.14}$	H	0.015
XTEJ1650–500	2001/2002	93.12 ± 1.26	$9.804^{+4.36}_{-3.56}$	$0.185^{+0.034}_{-0.052}$	H,I,S	0.016
XTEJ1859+226	1999/2000	$56.61^{+0.066}_{-0.084}$	$11.62^{+5.23}_{-4.27}$	$0.505^{+0.142}_{-0.093}$	H,I,S	0.18

The efficiency of angular-momentum (and mass) transport α , assuming no mass loss in the hot disk, and related quantities, sampled using our Bayesian hierarchical model, are presented.

The accretion state(s) reached in each outburst are indicated as hard (H), intermediate (I) or soft (S).

^a M_1 is unconstrained, so we assume $M_1 = 10M_\odot$.

^bDistance d is unconstrained, so we assume $d = 8$ kpc.

A large oxygen-dominated core from the seismic cartography of a pulsating white dwarf

N. Giammichele^{1,2}, S. Charpinet¹, G. Fontaine², P. Brassard², E. M. Green³, V. Van Grootel⁴, P. Bergeron², W. Zong^{1,5} & M. -A. Dupret⁴

White-dwarf stars are the end product of stellar evolution for most stars in the Universe¹. Their interiors bear the imprint of fundamental mechanisms that occur during stellar evolution^{2,3}. Moreover, they are important chronometers for dating galactic stellar populations, and their mergers with other white dwarfs now appear to be responsible for producing the type Ia supernovae that are used as standard cosmological candles⁴. However, the internal structure of white-dwarf stars—in particular their oxygen content and the stratification of their cores—is still poorly known, because of remaining uncertainties in the physics involved in stellar modelling codes^{5,6}. Here we report a measurement of the radial chemical stratification (of oxygen, carbon and helium) in the hydrogen-deficient white-dwarf star KIC08626021 (J192904.6+444708), independently of stellar-evolution calculations. We use archival data^{7,8} coupled with asteroseismic sounding techniques^{9,10} to determine the internal constitution of this star. We find that the oxygen content and extent of its core exceed the predictions of existing models of stellar evolution. The central homogeneous core has a mass of 0.45 solar masses, and is composed of about 86 per cent oxygen by mass. These values are respectively 40 per cent and 15 per cent greater than those expected from typical white-dwarf models. These findings challenge present theories of stellar evolution and their constitutive physics, and open up an avenue for calibrating white-dwarf cosmochronology¹¹.

Hydrogen-deficient DB-type white dwarfs are dying stars that have effective temperatures ranging from about 12,000 K to upwards of 35,000 K (ref. 12). They have undergone a late final flash as post-asymptotic-giant-branch (post-AGB) stars, which freed them from their thin remaining hydrogen envelope^{13–15}. Their deeper chemical stratification—as in any typical white dwarf—is otherwise determined by the rate of the nuclear reaction $^{12}\text{C}(\alpha, \gamma)^{16}\text{O}$, which synthesizes oxygen from the ashes of the core-helium-burning phase, by means of mixing processes associated with convection and rotation as well as the number of thermal pulses that occur in evolved AGB stars¹⁶. The descendant DB white dwarfs must bear the signature of such processes in their core, as well as the imprint of the still-ongoing settling of carbon and oxygen in the surrounding helium-rich envelope.

KIC08626021 (GALEX J192904.6+444708) is the first pulsating DB white dwarf (also known as the V777 Her class of variable stars) to be monitored extensively by the Kepler spacecraft for its pulsation properties⁸. This star shows an oscillation spectrum composed of non-radial g-modes, which are potentially sensitive to the deep interior of the star¹⁷. Using an analysis of 23 months of Kepler high-precision photometric data¹⁸, we exploit the eight well secured independent modes, which have periods ranging from 143.2 s to 376.1 s (Fig. 1, Extended Data Table 1 and Methods). These relatively short pulsation periods are—according to the period–luminosity relation¹⁷—typical of the blue (hot) edge of the V777 Her instability strip, as confirmed

by our spectroscopic determination of the atmospheric parameters T_{eff} (effective temperature; $29,360 \pm 780$ K) and $\log(g)$ (7.89 ± 0.05) (see Methods and Extended Data Fig. 1).

The seismic analysis performed here relies on a forward-modelling technique using parameterized static stellar models, which is independent of stellar-evolution calculations. This independence is crucial for providing an objective determination of the chemical stratification in the core of white-dwarf stars. Our analysis differs fundamentally from several previous attempts to provide asteroseismic constraints on such stars that rely on evolutionary models¹⁹ and therefore carry with them the many uncertainties of such models. Here, we take advantage of our most recent developments in the definition of parameterized white-dwarf models for asteroseismology^{9,10}. This method allows a flexible and exhaustive exploration of the structural configuration (core and envelope) of white dwarfs that outperforms all comparable approaches that have been attempted in the past. A historical and critical review of all of these developments has been presented recently⁹. In particular, none of these attempts has succeeded in reproducing accurately the observed periods and in securely extracting an independent measure of the core stratification. In a nutshell (see Methods for more details and Extended Data Fig. 2), we use a shape-optimization technique based on Akima splines to produce smooth, adjustable oxygen profiles in the core. This profile is optimized simultaneously with other parameters that define the star's full hydrostatic structure and the chemical stratification in its envelope. With this technique, we can ultimately recover, along with global parameters, the optimal model chemical stratification for the main constituents (here carbon, oxygen and helium) that is best able to reproduce the seismic observables, as hare-and-hounds experiments have demonstrated¹⁰.

When applied to KIC08626021, the seismic solution is uniquely determined around a well defined minimum of the fitted merit function in parameter space (see Methods, Extended Data Figs 3 and 4, and Extended Data Table 2). The optimal model is found to have $T_{\text{eff}} = 29,968 \pm 200$ K and $\log(g) = 7.917 \pm 0.009$, which match perfectly the independent measurements obtained from spectroscopy (within 0.8σ for the effective temperature and 0.5σ for the surface gravity). Other seismic inferences are provided in Table 1. The seismic values given in that table are derived from the optimal model solution, and the associated errors represent internal errors of the fit corrected with an estimation of external errors due to known potential systematics. These errors are evaluated from the likelihood function covering the full parameter space that is sampled during the optimization procedure²⁰ (Methods).

We find that the seismic fit perfectly reproduces the measured frequencies—that is, the seismic fit is well within the precision of the observations estimated at $\Delta\nu = 0.6$ nHz (or equivalently at $\Delta P = 38 \mu\text{s}$ for the period measurements) for this dataset (Extended Data Table 1). Note that, with nearly two years of Kepler data, the precision of the

¹Institut de Recherche en Astrophysique et Planétologie (IRAP), Université de Toulouse, Centre National de la Recherche Scientifique (CNRS), UPS, Centre National d'Études Spatiales (CNES), 14 Avenue Edouard Belin, F-31400 Toulouse, France. ²Département de Physique, Université de Montréal, Montréal, Québec H3C 3J7, Canada. ³Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, Arizona 85721, USA. ⁴Space sciences, Technologies and Astrophysics Research (STAR) Institute, Université de Liège, 19C Allée du six-août, B-4000 Liège, Belgium. ⁵Department of Astronomy, Beijing Normal University, Beijing 100875, China.

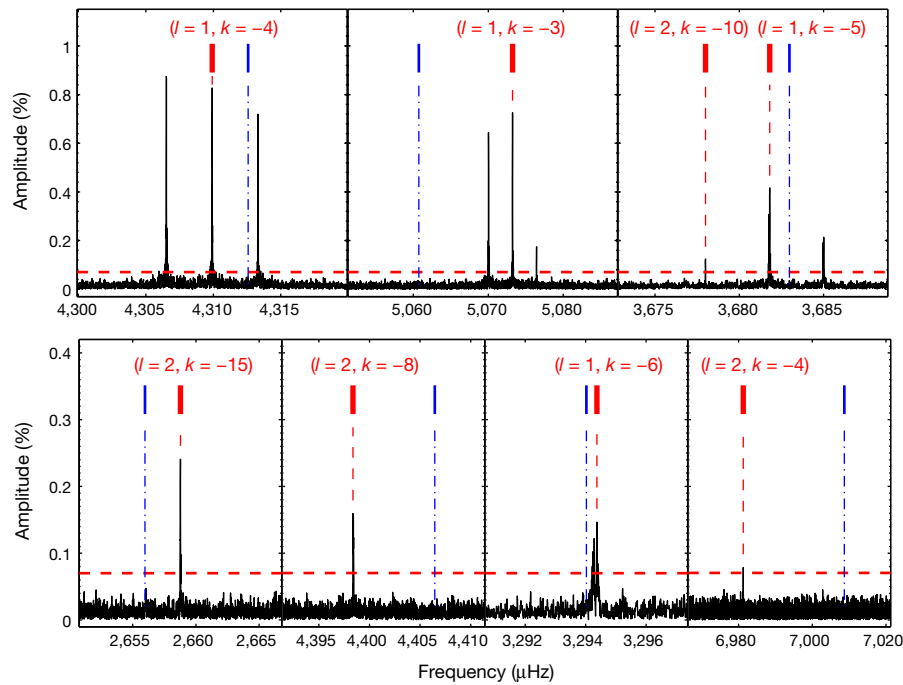


Figure 1 | Segments of the Lomb–Scargle periodogram for KIC08626021. This periodogram shows the pulsation amplitude (as a percentage) of the star’s mean brightness versus frequency around detected signals (in μHz), superimposed with the matched theoretical frequencies obtained from our optimal seismic model (red; eight gravity modes) and the latest published results for this star²¹ (blue; seven modes). Our asteroseismic modelling solution is able to reproduce the eight

main observed pulsation periods in KIC08626021 at the precision of the observations (see Extended Data Table 1). That solution provides the mode identification shown here in terms of the degree index l and radial order k . The other statistically significant peaks are part of frequency multiplets best explained in terms of rotational splitting (see Methods and Extended Data Fig. 6).

measured frequencies is extremely high. Consequently, this new seismic solution outperforms, by orders of magnitude, any former analysis, including the most recent result obtained for the same star, which could match the frequencies only to an average precision of $\Delta\nu = 7.8 \mu\text{Hz}$ (ref. 21; see Fig. 1). This major improvement in the global precision of the fit allows for a precise derivation of the internal stratification. The distribution of the oxygen mass fraction in the core of the star—which has remained mostly unconstrained thus far—can now be unambiguously derived, along with profiles for the other main chemical species (Fig. 2 and Extended Data Fig. 5).

We find that the central abundance of oxygen rises up to 86.2% ($\pm 4\%$), that is, 15% more than is predicted from evolutionary calculations². The extent of the homogeneous central part of the core is nearly doubled, reaching a fractional mass depth of $\log(q) \approx -0.7$ (or 0.45 solar masses, M_\odot), instead of $\log(q) \approx -0.35$ ($0.32 M_\odot$). According to standard evolutionary calculations, this homogeneous region derives from the former convective, helium-burning core, which is possibly enlarged by a non-negligible region (up to 25% in mass) that is composed of the ashes of a rather thick, helium-burning shell across a predicted partially mixed zone (the former semi-convective layers), and is ultimately fully homogenized owing to Rayleigh–Taylor instabilities²². Recently, however, models that match the average spacing of pulsation periods observed in core-helium-burning red-clump stars have favoured a treatment of the core boundary that is enlarged by substantial overshoot (the so-called maximal overshoot model); these models do not produce a partial mixing zone (there is no semi-convection)²³. In that situation, the oxygen plateau in the white-dwarf core may entirely be associated with the final extent of the fully mixed helium-burning core. This suggests that the extent of the progenitor convective core during the core-helium-burning phase should have encompassed masses of between $0.34 M_\odot$ and $0.45 M_\odot$ —larger by a factor of between 1.1 (in the case of the maximal semi-convection zone predicted) and 1.4 (in the absence of a semi-convection zone). Interestingly, similar trends involving a large convective

Table 1 | Derived properties of KIC08626021

Quantity	Estimated value
$\log[g \text{ (cm s}^{-2}\text{)}]$	7.92 ± 0.01
T_{eff} (K)	$29,968 \pm 198$
$X(\text{He})_{\text{env}}$	0.18 ± 0.04
$\log(q_1)$	-7.63 ± 0.2
$\log(q_2)$	-3.23 ± 0.1
$X(\text{O})_{\text{centre}}$	0.86 ± 0.04
$\log(q_3)$	-0.72 ± 0.03
$M(\text{He})/M_*$	$0.0113\% \pm 0.006\%$
$M(\text{C})/M_*$	$21.96\% \pm 4.2\%$
$M(\text{O})/M_*$	$78.03\% \pm 4.2\%$
M_*/M_\odot	0.570 ± 0.005
R_*/R_\odot	0.0138 ± 0.0001
L_*/L_\odot	0.137 ± 0.005
M_r^*	10.28 ± 0.03
d (pc) [†]	422 ± 45
P_{rot} (h)	46.3 ± 2.5
V_{eq} (km s ⁻¹)	0.36 ± 0.02
J (kg m ² s ⁻¹) [‡]	$(6.59 \pm 0.38) \times 10^{38}$
J/J_\odot	$1/291$
dP/dt_{197} (s s ⁻¹) [¶]	14.4×10^{-14} or 2.8×10^{-14}
dP/dt_{232} (s s ⁻¹) [¶]	15.1×10^{-14} or 2.8×10^{-14}
dP/dt_{271} (s s ⁻¹) [¶]	15.5×10^{-14} or 3.0×10^{-14}
t_{cool} (yr)	$(8.43 \pm 0.05) \times 10^6$

Along with primary quantities coming directly from our seismic analysis, several secondary properties are also derived as explained in Methods. $X(\text{He})$, mass fraction of helium; $\log(q_1)$, the fractional mass depth where the local value of the He mass fraction equals $(1 + X(\text{He})_{\text{env}})/2$ (where $X(\text{He})_{\text{env}}$ is the mass fraction of He in the mixed zone of the star’s envelope); $\log(q_2)$, the depth at which that mass fraction equals $X(\text{He})_{\text{env}}/2$; $M(\text{He})$, $M(\text{C})$ and $M(\text{O})$, mass of helium, carbon and oxygen, respectively; M_* , R_* and L_* , mass, radius and luminosity of the star, correspondingly; M_r , absolute magnitude in the r -band; d , distance to Earth (in parsecs); P_{rot} , rotation period; V_{eq} , equatorial velocity; J , total angular momentum; dP/dt_x , rate of period change of the mode with period x (in units of seconds per second), with (first number) or without (second number) neutrino emission; t_{cool} , cooling time of the star.

[‡]Based on a model atmosphere with the seismic values of $\log(g)$, T_{eff} and M_* , coupled to the uncertain value of the Sloan Digital Sky Survey magnitude $r = 18.40 \pm 0.20$ for KIC08626021.

[†]Assuming no reddening.

[‡]Assuming solid-body rotation throughout the star.

[¶]With and without neutrino cooling.

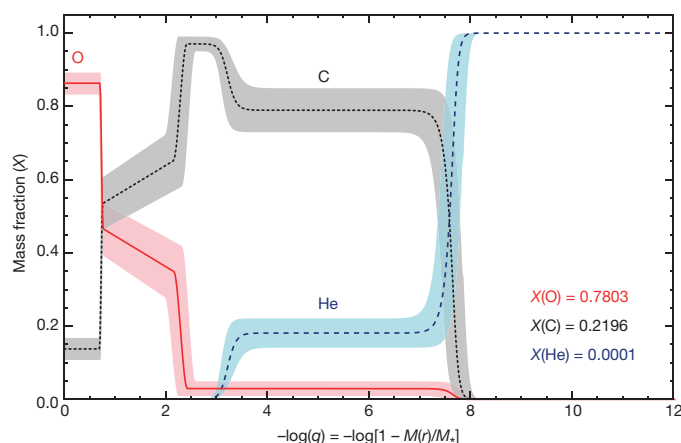


Figure 2 | Derived chemical stratification. The graph shows the distribution of oxygen (in red), carbon (in grey) and helium (in blue) obtained from the optimal seismic model. The estimated 1σ errors (shaded areas around each curve) are derived from the probability distributions calculated during the optimization process (Methods). The x axis shows the fractional mass depth (with $\log(q) = 0$ corresponding to the centre of the star). The total mass fractions for each element are indicated at the bottom right.

core were disclosed in a study of pulsating hot B subdwarf (sdB) stars^{24–26}, in particular because in these cases the probed cores may not yet have reached their maximal extent. These results will allow researchers to exclude mixing schemes that cannot produce enough oxygen or extended cores, and guide a better understanding of mixing processes.

The behaviour of the two outward descents in the oxygen mass stratification that bear the imprint of helium-shell-burning processes from earlier stages of evolution does not differ notably from the predictions of standard evolutionary calculations. The total oxygen content of the white-dwarf core reaches $78.0\% \pm 4.2\%$, much higher than the expected value of around 64% for a standard evolutionary white-dwarf model of the same mass². Remarkably, we note the absence in our solution of a triple transition in which carbon, oxygen and helium coexist—a feature usually expected from evolution calculations. Although puzzling at first sight, this issue could be explored in relation to the number of thermal pulses that occur to erode the bottom of the helium envelope during post-AGB evolution³, particularly given that KIC08626021's envelope is found to be quite thin. Finally, our finding of a double-layered helium envelope fits with the presently accepted scenario involving ongoing settling of helium after the late thermal pulse that produced this DB white dwarf. Overall, these new seismic constraints offer opportunities for testing stellar-evolution models and their constitutive physics (particularly processes that include convection, overshooting, semi-convection, and nuclear physics). The next step will be to explore late stages of stellar evolution, with the goal of reproducing the seismically derived chemical stratification of the white-dwarf core.

The chemical profile derived here from asteroseismology, applied to computations of white-dwarf cooling ages, leads to an estimate of $(8.43 \pm 0.05) \times 10^6$ yr for the cooling age of KIC08626021, independently of its pre-white-dwarf history. By comparison, uncertainties in the rates of nuclear reactions in the pre-white-dwarf phases alone lead to an uncertainty on that age that is ten times larger⁵ than the uncertainty calculated above. Hence, the asteroseismological approach described here provides a powerful calibration of the internal composition profile of white dwarfs, with direct benefit to cosmochronology.

Finally, the total carbon/oxygen ratio found here for KIC08626021—which is lower than expected—has implications for studies of type Ia supernovae²⁷. A smaller carbon/oxygen ratio leads to less ^{56}Ni being produced during the supernova explosion. This has a direct impact on the light curve of a type Ia supernova (which is powered by the decay

of ^{56}Ni), with repercussions for cosmology, where these supernovae are widely used as standard candles and probes—notably for determining the cosmological equation of state.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 9 June; accepted 7 November 2017.

Published online 8 January 2018.

- Fontaine, G. & Wesemael, F. in *Encyclopedia of Astronomy and Astrophysics 1894–1901 (IOP, 2000)*.
- Salaris, M., Cassisi, S., Pietrinferni, A., Kowalski, P. M. & Isern, J. A large stellar evolution database for population synthesis studies. VI. White dwarf cooling sequences. *Astrophys. J.* **716**, 1241–1251 (2010).
- De Gerónimo, F. C., Althaus, L. G., Corsico, A. H., Romero, A. D. & Kepler, S. O. Asteroseismology of ZZ Ceti stars with fully evolutionary white dwarf models. I. The impact of the uncertainties from prior evolution on the period spectrum. *Astron. Astrophys.* **599**, A21–A28 (2017).
- Fontaine, G., Brassard, P. & Bergeron, P. The potential of white dwarf cosmochronology. *Publ. Astron. Soc. Pac.* **113**, 409–435 (2001).
- Fields, C. E., Farmer, R., Petermann, I., Iliadis, C. & Timmes, F. X. Properties of carbon-oxygen white dwarfs from Monte Carlo stellar models. *Astrophys. J.* **823**, 46 (2016).
- Salaris, M. White dwarf cosmochronology: techniques and uncertainties. *Proc. Int. Astron. Soc.* **258**, 287–298 (2009).
- Gilliland, R. L. et al. Kepler asteroseismology program: introduction and first results. *Publ. Astron. Soc. Pac.* **122**, 131–143 (2010).
- Østensen, R. H. et al. At last—a V777 Her pulsator in the Kepler field. *Astrophys. J.* **736**, L39–L44 (2011).
- Giammichele, N., Charpinet, S., Fontaine, G. & Brassard, P. Toward high-precision seismic studies of white dwarf stars: parametrization of the core and tests of accuracy. *Astrophys. J.* **834**, 136–162 (2017).
- Giammichele, N., Charpinet, S., Brassard, P. & Fontaine, G. The potential of asteroseismology for probing the core chemical stratification in white dwarf stars. *Astron. Astrophys.* **598**, A109–A116 (2017).
- Tremblay, P.-E., Kalirai, J. S., Soderblom, D. R., Cignoni, M. & Cummings, J. White dwarf cosmochronology in the solar neighborhood. *Astrophys. J.* **791**, 92–100 (2014).
- Bergeron, P. et al. A comprehensive spectroscopic analysis of DB white dwarfs. *Astrophys. J.* **737**, 28–51 (2011).
- Iben, I., Jr, Kaler, J. B., Truran, J. W. & Renzini, A. On the evolution of those nuclei of planetary nebulae that experience a final helium shell flash. *Astrophys. J.* **264**, 605–612 (1983).
- Herwig, F., Blocker, T., Langer, N. & Driebe, T. On the formation of hydrogen-deficient post-AGB stars. *Astron. Astrophys.* **349**, L5–L8 (1999).
- Miller Bertolami, M. M., Althaus, L. G., Serenelli, A. M. & Panei, J. A. New evolutionary calculations for the born again scenario. *Astron. Astrophys.* **449**, 313–326 (2006).
- Straniero, O., Dominguez, I., Imbriani, G. & Piersanti, L. The chemical composition of white dwarfs as a test of convective efficiency during core helium burning. *Astrophys. J.* **583**, 878–884 (2003).
- Fontaine, G. & Brassard, P. The pulsating white dwarf stars. *Publ. Astron. Soc. Pac.* **120**, 1043–1096 (2008).
- Zong, W., Charpinet, S., Vauclair, G., Giammichele, N. & Van Grootel, V. Amplitude and frequency variations of oscillation modes in the pulsating DB white dwarf star KIC08626021. The likely signature of nonlinear resonant mode coupling. *Astron. Astrophys.* **585** (A22), 1–14 (2016).
- Romero, A. D. et al. Toward ensemble asteroseismology of ZZ Ceti stars with fully evolutionary models. *Mon. Not. R. Astron. Soc.* **420**, 1462–1480 (2012).
- Giammichele, N., Fontaine, G., Brassard, P. & Charpinet, S. A new analysis of the two classical ZZ Ceti white dwarfs GD 165 and Ross 548. II. Seismic modeling. *Astrophys. J. Suppl. Ser.* **223**, 10–36 (2016).
- Bischoff-Kim, A., Østensen, R. H., Hermes, J. J. & Provencal, J. Seven-period asteroseismic fit of the Kepler DBV. *Astrophys. J.* **794**, 39–46 (2014).
- Salaris, M. et al. The cooling of CO white dwarfs: influence of the internal chemical distribution. *Astrophys. J.* **486**, 413–419 (1997).
- Constantino, T., Campbell, S. W., Christensen-Dalsgaard, J., Lattanzio, C. & Stello, D. The treatment of mixing in core helium burning models—I. Implications for asteroseismology. *Mon. Not. R. Astron. Soc.* **452**, 123–145 (2015).
- Van Grootel, V. et al. Early asteroseismic results from *Kepler*: structural and core parameters of the hot B subdwarf KPD 1943+4058 as inferred from g-mode oscillations. *Astrophys. J.* **718**, L97–L101 (2010).
- Van Grootel, V., Charpinet, S., Fontaine, G., Green, E. M. & Brassard, P. Structural and core parameters of the hot B subdwarf KPD 0629–0016 from CoRoT g-mode asteroseismology. *Astron. Astrophys.* **524**, A63 (2010).
- Charpinet, S. et al. Deep asteroseismic sounding of the compact hot B subdwarf pulsator KIC02697388 from *Kepler* time series photometry. *Astron. Astrophys.* **530**, A3 (2011).
- Dominguez, I., Höflich, P. & Straniero, O. Constraints on the progenitors of type Ia supernovae and implications for the cosmological equation of state. *Astrophys. J.* **557**, 279–291 (2001).

Acknowledgements S.C., N.G. and W.Z. acknowledge financial support from Programme National de Physique Stellaire (PNPS) of CNRS/INSU, France, and from the Centre National d'Études Spatiales (CNES, France). We also acknowledge support from the Agence Nationale de la Recherche (ANR, France) under grant ANR-17-CE31-0018, funding the INSIDE project. This work was granted access to the high-performance-computing resources of the CALMIP computing centre under allocation number 2017-p0205. This work was supported by the Fonds Québécois de la Recherche sur la Nature et les Technologies (FORNT, Canada) through a postdoctoral fellowship awarded to N.G. G.F. also acknowledges the contribution of the Canada Research Chair Program, and W.Z. the LAMOST fellowship as a young researcher, supported by the Special Funding for Advanced Users, budgeted and administrated by the Center for Astronomical Mega-Science, Chinese Academy of Sciences. V.V.G. is an F.R.S.-FNRS Research Associate. The authors acknowledge the Kepler team and everyone who has contributed to making this mission possible. Funding for the Kepler mission is provided by NASA's Science Mission Directorate.

Author Contributions N.G. wrote the manuscript and performed the seismic analysis of KIC08626021. S.C. and G.F. contributed to the writing of the manuscript. S.C., P.Br. and N.G. contributed to the development of the numerical codes used in this analysis. E.M.G. obtained the spectroscopic data for KIC08626021. P.Be. and G.F. performed the model atmospheric analysis. W.Z. provided Fig. 1. V.V.G., G.F. and M.A.D. performed non-adiabatic complementary analysis. All authors discussed the results and contributed to their interpretation.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to N.G. (noemi.giammichele@irap.omp.eu).

Reviewer Information *Nature* thanks M. Salaris, O. Straniero and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Our method revolves around minimizing a merit function defined by the sum of the squared differences between theoretical and observed periods, in order to isolate an optimal self-consistent seismic model that best reproduces the oscillation properties of the star under consideration²⁰. The method and tools used here are essentially the same as those used in the recent successful test²⁰ in the case of the pulsating ZZ Ceti stars GD 165 and Ross 548, except that we include the detailed parameterization of the chemical composition profile in the star's core⁹. The method is inspired by work on pulsating hot subdwarf stars²⁸, which has been developed further over the years^{29–31}. Most recently, a discussion of our continuing quest for an objective approach to quantitative asteroseismology has been presented with application to white-dwarf stars in particular^{10,32}. The forward approach used here consists of comparing the pulsation periods of large numbers of theoretical stellar models—defined by N free parameters—to the observed periods of the pulsating star of interest. By optimizing this process, it is possible to infer the structural parameters of the appropriate stellar model via the best fit to the observed period spectrum. Given that we have no preconceived idea of the mode identification (except for the fact that the observed periods are likely to have a degree index, l , of 1 or 2, in keeping with what is known about identified gravity modes in white dwarfs), the optimization process has to be twofold. The technique thus relies on a double-optimization scheme that, first, best matches the observed periods with the periods calculated from a theoretical spherical stellar model, using a χ^2 -type formalism viewed as a combinatorial optimization problem. The first optimization therefore consists of finding the combination (or mode identification) that leads to the best possible simultaneous match of all the observed periods for that given model. It is based on a stochastic method using an evolutionary approach to optimization problems³². The method allows for the automatic and objective identification of modes, without any prior knowledge of the mode identification and outputs from the best fit, the degree l and the radial order k for each mode. The second optimization uses a multimodal optimizer based on a massively parallel hybrid genetic algorithm to find the best-fitting model in the N -dimensional parameter space, the quality of the fit being quantified by a merit function, S^2 (an unweighted χ^2):

$$S^2(a_1, a_2, \dots, a_N) = \sum_{i=1}^{N_{\text{obs}}} (P_{\text{obs}}^{(i)} - P_{\text{th}}^{(i)})^2 \quad (1)$$

where N_{obs} is the number of observed periods; a_1, a_2, \dots, a_N are the parameters of the model; and each $P_{\text{obs}}^{(i)}, P_{\text{th}}^{(i)}$ is a pair of matched observed–theoretical periods for this model. The code used, named LUCY³⁰, performs a complete and exhaustive exploration of the entire model parameter space and gives feedback on the uniqueness (or not) of the solution. Some illustrations of the capabilities of the code and further details can be found in ref. 32.

Parameterized models. The evaluation of S^2 requires computation of a large number of theoretical stellar models from which we obtain the period spectra. This is best achieved through the use of static, parameterized equilibrium models²⁰. Note that these are full stellar models (not envelope models), and that, for a cooling white dwarf, the well calibrated relation $L(r) \propto M(r)$ ensures a good description of the luminosity profile. To define a full, static V777 Her white-dwarf model, it is necessary to specify the surface gravity (or the mass via the mass–radius relation; we prefer to use the surface gravity because it allows direct comparison with atmospheric parameters derived from spectroscopy), the effective temperature, the envelope layering, the core composition and the convective efficiency. In the case of low-order, short-period g -modes such as those observed in KIC08626021, the periods are quite insensitive to the precise choice of the latter parameter (see, for example, figure 42 in ref. 17), so we fixed it to the ‘flavour’ ML2/ $\alpha = 1.25$, as calibrated through DB spectroscopy³³.

For a DB star, the envelope layering must be defined by a minimum of five parameters. This is illustrated in Extended Data Fig. 2, which shows the general shape to be expected for the helium profile in such a star. The envelope is characterized by two zones: first, an upper pure-helium layer with a thickness specified by $\log(q_1)$, produced by the gravitational settling of carbon and oxygen in earlier evolutionary phases; and second, a lower mixed zone with a base located at $\log(q_2)$ and in which separation between helium, carbon and oxygen still goes on. The quantity $X(\text{He})_{\text{env}}$ indicates the helium mass fraction in the mixed zone and represents the atmospheric helium content in the presumed PG 1159 ancestor. These quantities are related in such a way that $\log(q_1)$ corresponds to the fractional mass depth at which the local value of the helium mass fraction is equal to $(1 + X(\text{He})_{\text{env}})/2$, while $\log(q_2)$ corresponds to the depth where that mass fraction is equal to $X(\text{He})_{\text{env}}/2$. In addition to these three parameters, the stratification of helium in the envelope of a DB white dwarf must also be characterized by the actual shape of the profile in the composition transition zone between the pure-helium outermost layer and the flat portion of the curve

in the mixed zone, as well as the shape of the corresponding transition profile between the mixed zone and the core, where $X(\text{He})$ goes to zero.

To handle this problem, we first compute, for each transition zone, the actual profile obtained from the strict condition of diffusive equilibrium between ordinary diffusion and gravitational settling (we neglect the contribution of thermal diffusion here), which depends on the local conditions of density, temperature, and average charge of each ionic species of interest. Recognizing next that diffusive equilibrium cannot have been reached over the cooling lifetimes of DB white dwarfs, especially in the deeper regions of the star, we introduce a free parameter that acts as a multiplicative factor applied to the equilibrium profile. Varying that parameter—named P_1 for the upper transition zone and P_2 for the lower one—leads to a profile that may be steeper or shallower than the equilibrium one. The actual functional form used in this approach belongs to the so-called family of sigmoid functions. The implicit expression governing the chemical profile in such a transition zone is given by equation (1) of ref. 9. The parameters P_1 and P_2 have no particular physical meaning themselves. To summarize this, the envelope layering in a DB white dwarf model is defined by the specification of $X(\text{He})_{\text{env}}, \log(q_1), \log(q_2), P_1$ and P_2 .

To describe the complicated chemical stratification expected in the core of a carbon–oxygen white dwarf, we adopt a seven-parameter description⁹. This follows from the realization that g -modes can sometimes be confined in the deep core of a cool white dwarf²⁰ (this was the case for Ross 548, but not for GD 165), and thus could serve as probes of that core. In ref. 20, it was crudely assumed for simplicity that the carbon–oxygen core is homogeneous in composition, so that it could be characterized by a single free parameter, the carbon/oxygen ratio. Although the chemical composition of the core of a carbon–oxygen white dwarf is not expected to be homogeneous according to stellar evolution theory, this simple approach at least leads to an estimate of the bulk composition in a case, such as that of Ross 548, where there are detected modes with non-negligible amplitudes and weight functions extending into the core. The lesson here is that we should be prepared *a priori* to exploit potentially confined modes of this sort.

Figure 1 of ref. 9 illustrates in a schematic way the proposed parameterization for the carbon–oxygen core of a typical white dwarf. Strictly speaking, that parameterization applies to the cores of hydrogen–atmosphere, or DA, white dwarfs. For a DB star, an eighth parameter should be added; that is, the value of $X_{\text{env}}(\text{O})$ at the outer boundary of the core should not be set to zero as is the case for a DA star, but to some value, variable from one object to another, that reflects the atmospheric oxygen content in the previous PG 1159 evolutionary phase. This atmospheric pollution is the direct consequence of the violent mixing event that is associated with the born-again scenario at the origin of helium–atmosphere white dwarfs. As discussed in detail in ref. 9, eight control points are necessary to fully define a two-transition chemical profile in the core of a DB white dwarf. These control points correspond to the following parameters: core O, $t1$, $\Delta t1$, $t1(\text{O})$, $t2$, $\Delta t2$, $t2(\text{O})$ and $X_{\text{env}}(\text{O})$ (see figure 1 of ref. 9). The flexibility of our parameterization allows us to also test different structural configurations during the same process; for example, we can obtain and test a single-layered envelope when $q_1 = q_2$, or a single-transition chemical profile in the core by only using the subset of core O, $t1$ and $\Delta t1$, and by setting $t1(\text{O})$ and $t2(\text{O})$ to 0.

In the end, we carry out our search in parameter space by optimizing a total of 15 parameters: the surface gravity, the effective temperature, the five quantities defining the helium-rich envelope, and the eight parameters characterizing the carbon–oxygen core. Once the structure is fully defined, the following step is to evaluate the adiabatic pulsation properties of the model—assumed to be purely spherical—using an efficient and robust code based on finite-element techniques^{34,35}.

Pulsation properties. To carry out our seismic study, we use a recent thorough analysis¹⁸, which has shed new light on the 23 months of Kepler photometric data gathered for KIC08626021. In particular, even if the 13 retained frequencies are consistent with previous analyses²¹, interpretation of one of the structures leads to an important correction. The structure in the 3,677–3,686 μHz range, identified previously as a triplet^{21,36}, is actually a doublet, f_3 (ref. 18), with an independent mode, f_7 , close by. The seismic modelling is greatly influenced by the addition of this eighth low-order frequency, as it must probe a somewhat different part of the star than the other modes, thus adding to the total information content available from the observations and contributing to a more detailed seismic view of KIC08626021. The list of extracted periods (frequencies) and their specifications are provided in table 1 of ref. 18; we refer the reader to this work for details of the extraction process and the frequencies detected. We use the same convention for the sake of simplicity.

Our seismic analysis is thus based on the following eight frequencies: $f_1, f_2, f_3, f_4, f_5, f_6, f_7$ and f_9 . There is fine structure around three of these peaks in the Fourier domain: f_1 and f_2 are two clearly identified and nearly symmetric triplets in frequency space, while f_3 shows a doublet structure with components separated by a spacing comparable to those seen in the triplets. Interpreted as rotational splitting, this fine structure corresponds to a rotation timescale of about 42.5 h

if the three modes are assumed to be dipole modes and if the asymptotic expression for the first-order solid-body rotation coefficient, C_{kl} , is used: that is, $C_{kl} = 1/(l^2 + 1) = 0.5$. Out of the remaining four periodicities listed in table 1 of ref. 18, f_8 and f_{10} are linear-combination frequencies of two of the eight basic modes, and as such, they are excluded from our analysis. In addition, we do not consider f_{11} and f_{12} because they do not reach the secure threshold of 5.6σ (ref. 18).

The optimal model. The search for best-fitting solutions starts with the largest possible parameter space for DBV stars. Constraints on T_{eff} and $\log(g)$ rely on typical spectroscopic values found for DBV stars, being as inclusive as possible. Evolutionary tracks² influenced the search ranges for the eight parameters defining the C–O core: core O, t_1 , Δt_1 , $t_1(\text{O})$, t_2 , Δt_2 , $t_2(\text{O})$ and $X_{\text{env}}(\text{O})$. We consider all modes of degree $l = 1$ and $l = 2$ in the period range from 100 s to 500 s, which encompasses the range spanned by the eight main observed modes in KIC08626021. Modes with higher degree are excluded, as we expect them to be hardly visible owing to cancellation effects of their surface geometry. We thus searched parameter space in 15 dimensions in order to best fit eight periods. *A priori*, this seems like an underconstrained problem, but this simplistic point of view is in fact incorrect in the context of the present optimization exercise—a highly nonlinear problem. This point has been discussed at length^{9,10}.

Keeping this in mind, we singled out a model in parameter space that can match the observed periods at an unprecedented level of precision, defined by a merit function of $S^2 = 1.4 \times 10^{-15}$. Information on the best-fitting model—including details of the period spectrum and mode identification—is presented in Extended Data Table 1. This very precise seismic match reaches (for the first time, to our knowledge) the actual observational limit. In this case, the observational limit is $\Delta P = 0.000038$ s or $\Delta\nu = 0.00060 \mu\text{Hz}$, based on the two years of Kepler data on KIC 08626021—justifying the number of digits given in Extended Data Table 1 for these quantities.

Extended Data Fig. 3 shows a projection map of the merit function onto the $T_{\text{eff}}\text{--}\log(g)$ plane. The merit function is normalized in such a way that the global minimum is equal to one, on a logarithmic scale. As can be seen, the seismic effective temperature falls well within the spectroscopic 1σ box. The seismic value of $\log(g)$ is also in excellent agreement with the spectroscopic determination to within 1σ . The actual seismically inferred values of T_{eff} and $\log(g)$, along with the other defining parameters for this unique solution, are listed in Extended Data Table 2.

The pure He layer is rather thin with a value of $\log(q_1) = -7.63$, representing the position of the transition between the mixed-composition envelope and the pure He layer in the $\log(q)$ coordinate. The envelope composed of a mixture of oxygen, carbon and helium, with a global mass fraction of 18.1% He, is defined by $\log(q_2) = -3.23$. This leads to a total helium mass fraction (integrated over the full model) of $\log[M(\text{He})/M_*] = -3.948$ or, equivalently, $M(\text{He})/M_* = 0.0113\%$. Using the global and shape parameters listed in Extended Data Table 2, we report our primary results in Table 1 (these results are the top 11 entries in the table and include the total mass, which is a direct product of the constitutive physics used in our stellar models). From those, one can infer interesting secondary quantities such as the radius, the luminosity, the absolute magnitude in the Sloan Digital Sky Survey (SDSS) r -band, and an estimate of the distance to KIC08626021. Other secondary properties, reported in the lower part of Table 1, can also be inferred as discussed below.

An integral part of our approach is the requirement that the neighbourhood of the optimal model in parameter space be thoroughly explored in order to assess the statistical significance of the solution and to provide estimates of uncertainties on each derived parameter. As an example, Extended Data Fig. 4 shows the histograms obtained for the probability distribution of six interesting parameters used during the present optimization. The technique used relies on the likelihood function calculated from the sampling of the merit function S^2 during the search for the best-fitting model. This method³¹ allows us to make more quantitative statements on the value of each parameter by statistically estimating that value and its associated errors from the seismic fit²⁰. We also incorporate in the error budget, at least partially, an evaluation of external errors resulting from potential systematics. In a nutshell, all of the histograms show narrow peaks, with the mean and median values of the distribution being very close to each other. The derived uncertainties are listed in Table 1.

The inferred chemical stratification. Our central result is the unravelling of the complex internal chemical stratification of KIC08626021 (see Fig. 2). The 1σ uncertainties in Fig. 2 (the shaded areas) are derived from the probability distributions calculated during the optimization process. For completeness, we show, in Extended Data Fig. 5, the individual normalized probability distributions for oxygen, carbon and helium obtained during this exercise.

We also point out that our ability to probe the core composition is intimately associated with the presence of detected modes with amplitudes and weight functions sufficiently important in the core for the pulsation periods to show a sensitivity to the local chemical composition and stratification²⁰. Extended Data

Fig. 6a illustrates that all of the eight modes of interest in our optimal seismic model (those identified with the eight observed pulsations) have at least some sensitivity to the core structure, as measured by the non-negligible values of their weight functions at those depths. Some of the modes, such as those with $l = 1$, $k = -3$ and $l = 2$, $k = -4$ (that is, the confined modes), are particularly good probes of the core composition²⁰.

In comparison with the much cooler ($T_{\text{eff}} \approx 12,000$ K) ZZ Ceti stars²⁰ (see figure 6 of ref. 20 in particular), we find that the capacity of seismic waves to sound the core structure of a hot V777 Her pulsator such as KIC08626021 is much higher. This is explained by the well known effect of amplitude-mode migration associated with increased overall degeneracy in the cooler ZZ Ceti stars, which pushes outwards, and away from the core, the non-negligible parts of the weight function (see, for example, figure 8 of ref. 17).

Comparison with previous studies. It is also appropriate to compare our seismic results with those obtained in previous (adiabatic) studies dedicated to KIC08626021. Following the report of the discovery of five distinct pulsation modes in this star on the basis of one month of Kepler photometry⁸, three different groups presented preliminary seismic analyses. The first study³⁷ made use of parameterized evolutionary models computed with the white dwarf evolution code (WDEC). Five defining quantities were varied: the effective temperature, the total mass, two parameters characterizing the core, and one free parameter describing the envelope. Following this, an independent seismic study³⁸ was carried out based on evolutionary models with a full history from the zero-age main-sequence (ZAMS). In principle, such models are the most realistic ones available and are defined only by the total mass and initial chemical composition.

However, by design, these models show no flexibility in how they treat a star's constitutive physics, and they hide defects, including accumulated numerical noise and artefacts (see ref. 20 for the need for parameterized models in white-dwarf seismology). The third investigation³⁹ used an earlier version of the approach we use here, with the big difference that there was no detailed core parameterization. The core was assumed to be homogeneous in composition and characterized by a single quantity, the carbon-to-oxygen ratio. Specifying the effective temperature, the surface gravity, and a three-parameter description of the envelope, these authors³⁹ searched a six-dimensional parameter space using full, but static models. They obtained two solutions, the first assuming that all five modes are dipole modes, and the second allowing the possibility that the modes belong to degree index $l = 1$ or 2. Finally, the addition of two newly uncovered frequencies from the extended dataset³⁶ motivated another attempt at deciphering KIC08626021 (ref. 21), again with the use of parameterized evolutionary models computed with WDEC. In this case, an additional parameter (making a total of six) was included in the search in parameter space, in order to provide a more flexible description of the envelope. Two possible solutions were presented²¹.

As a first step, we contrast the mode identification and the normalized merit function of each of the previous studies with our work in Extended Data Table 3. The normalized merit function, s^2 , is defined as the merit function presented in the equation above, divided by the number of fitted periods. It allows us to put all seismic studies on KIC08626021 on the same footing for purposes of comparison. Note that our work is based on reanalysis of the nearly two years of Kepler data, which added an extra frequency¹⁸ to the frequency spectrum used by the most recent study²¹. We also note that to fit simultaneously an increased number of observed periods is a more demanding challenge, especially when they are low-order modes, because they contain extra information about the pulsating star and are sensitive to any small change in the modelled structure of the star. The resulting parameter space becomes more intricate and complex to resolve.

From Extended Data Table 3, we find it reassuring that the mode identification appears to be quite robust. Indeed, except for the case of the mode with a period of 376.11 s, all of the different seismic analyses lead to the same values of l and k for the observed periods. In the former case, the proposed identification has been either $l = 1$ and $k = 8$, or $l = 2$ and $k = 15$ (which are our derived values). Note that the identification of $l = 1$ and $k = 8$ for the 376.11 s pulsation in solution 1 of ref. 39 is 'artificial' in the sense that only dipole modes were considered in the period-matching exercise. By letting the five pulsations available free to be $l = 1$ or $l = 2$ modes, these authors obtained a much improved fit, including the identification of $l = 2$ and $k = 15$ for the 376.11 s period (their solution 2).

In terms of the quality of the fit, as measured by the normalized merit function, Extended Data Table 3 reveals a clear correlation between the method used and the results obtained. Considering first the three initial analyses based on the detection of five pulsations, we find that, first, the standard method of using detailed evolutionary models³⁸ gives a poor fit to the observed periods ($s^2 = 3.74$); second, a much improved fit ($s^2 = 7.84 \times 10^{-2}$) is obtained when using parameterized evolutionary models³⁷; and third, a further improved fit ($s^2 = 2.51 \times 10^{-5}$) is obtained when using parameterized static models³⁹. We believe that this 'progression' is notable in that the latter method is the most optimized one for searching parameter space in

fine detail. While solution 2 of ref. 39 shows a good fit to the observed periods, the computed periods still do not match the precision of the observations. On can conclude from this that a complicated object such as KIC08626021 cannot be modelled precisely in terms of six parameters only. By comparison, we have been able to fit essentially perfectly ($\chi^2 = 1.75 \times 10^{-16}$) the eight periods now detected in that star. This orders-of-magnitude improvement is the direct result of the flexibility of our new parameterization, coupled with an efficient algorithm that finds the optimal solution and evaluates finely the shape of the merit function in parameter space.

Extended Data Table 4 lists the main parameters defining the structure of our white-dwarf model compared with results from previous studies. Note that $\log(q_1)$ corresponds to M_{He} , $\log(q_2)$ corresponds to M_{env} , and $\log(q_3)$ is directly linked to q_{im} (refs 21, 37). These two studies make use of the same parameterization, with the only difference being that M_{env} was kept fixed during their first attempted seismic analysis. The chemical structure found for the core and envelope³⁸ is kept fixed according to their predictions from evolutionary computations, but they give the values of these three reference points for comparison purposes.

Inferences of the global parameters T_{eff} and $\log(g)$ (mass) are all within 1σ of the spectroscopic values, except for a specific analysis³⁸. Reference points from the He-envelope layering— $\log(q_1)$ and $\log(q_2)$ —are also similar in values. The main differences arise from the core definition $\log(q_3)$ and $X(\text{O})_{\text{centr}}$, where our solution presents an extended central carbon–oxygen core as well as a higher composition in oxygen in this homogeneous part.

Spectroscopic constraints. The first analysis of the time-averaged optical spectrum of KIC08626021 revealed the DB nature⁸ of this relatively faint star at a Kepler magnitude, K_p , of 18.46. These authors used a low signal-to-noise ratio (S/N) classification spectrum gathered with the intermediate-dispersion spectrograph and imaging system (ISIS) on the William Herschel Telescope, combined with spectroscopic model grids⁴⁰. The mixing-length (ML) prescription of $ML2/\alpha = 1.25$ for DB models³³ was adopted. The standard fitting procedure⁴¹ gave $T_{\text{eff}} = 24,900 \pm 750$ K and $\log(g) = 7.91 \pm 0.07$. Of these two preliminary estimates, that of the effective temperature revealed itself somewhat suspect, given that KIC08626021 belongs to the shortest-period pulsators of all of the known V777 Her stars and, therefore, has to be among the hottest of the class according to the well known period–luminosity relation depicted, for example, in figure 25 of ref. 17. Hence, the true effective temperature of KIC08626021 has to be much higher than the estimate provided⁸, which is more characteristic of the red edge of the V777 Her instability strip⁴². The need for a better optical spectrum of KIC08626021 was subsequently emphasized³⁷ because of the apparent conflict between the initial spectroscopic estimate of the effective temperature and the seismic model, which indicated a value of $T_{\text{eff}} \approx 29,200$ K.

Shortly after the variability of KIC08626021 was reported, we obtained a better spectrum with the help of the Boller and Chivens Cassegrain spectrograph attached to the Steward Observatory's 2.3-m Bok Telescope on Kitt Peak. Seven individual spectra, each with an exposure time of 1,800 s, were gathered on 26 June 2011 (universal time). We used the 400 mm^{-1} first-order grating with a $2.5''$ slit to obtain a spectrum with a typical resolution of about 8.7 \AA over the wavelength interval 3,620–6,900 \AA . The instrument rotator was set before each exposure, in order to align the slit within about 2° of the parallactic angle at the midpoint of the exposure. Helium–argon comparison spectra were taken immediately after the stellar exposure. The spectra were bias-subtracted, flat-fielded, background-subtracted, optimally extracted, wavelength-calibrated, and flux-calibrated using standard image reduction and analysis facility (IRAF) tasks. They were then shifted to the same relative velocity before being combined into a single spectrum (although the velocity differences between the seven spectra were unsubstantial). The total exposure time was 12,600 s, much larger than the periods of the detected pulsations (which all have low amplitudes of less than 1% of the mean brightness of the star), thus ensuring that the combined spectrum is representative of the mean effective temperature and surface gravity. The combined spectrum has a mean value for S/N of around 50, which is excellent for a faint star such as KIC08626021.

Two of us presented a preliminary analysis of that time-averaged spectrum³⁹. More recently, using the same technique¹², we repeated the spectroscopic analysis. Our updated fit is presented in Extended Data Fig. 1. The fit to the available lines is quite good, and leads to the following atmospheric parameters: $T_{\text{eff}} = 29,360 \pm 780$ K, $\log(g) = 7.89 \pm 0.05$, and $\log(\text{H}/\text{He}) = -3.0 \pm 0.1$. Interestingly, we detect a small trace of hydrogen, making KIC08626021 one of the hottest known DBA white dwarfs. At the same time, and more importantly in this context, our revised estimate of the effective temperature also makes KIC08626021 the second hottest of the known V777 Her pulsators (as it should be on the basis of its short observed periods). Pulsation was confirmed⁴³ for the known DB star PG 0112+104, from the K2 dataset, making it the hottest known V777 Her pulsator, with a most recent estimate of $T_{\text{eff}} = 31,040 \pm 1,060$ K (ref. 12). In our seismic analysis, the spectroscopic estimates of T_{eff} and $\log(g)$ are used as important independent constraints. Any seismic model to be found must be compatible

with these constraints in order to be credible. We note in this context that our spectroscopic analysis is perfectly compatible with an independent study carried out⁴⁴ on the basis of Cosmic Origins Spectrograph (COS)/Hubble Space Telescope (HST) spectroscopy and of SPY optical data. These authors infer the values of $T_{\text{eff}} = 30,000 \pm 1,000$ K and $\log(g) = 7.89 \pm 0.15$ for KIC08626021, solidifying the spectroscopic constraints used here.

Internal rotation profile. Given a seismic model, it is possible to exploit the fine structure uncovered by Kepler in three of the eight main modes detected in KIC08626021 (ref. 18). Interpreted as rotational splitting, the triplet structures seen in the $l = 1, k = -3$ (197.1 s) and in the $l = 1, k = -4$ (232.0 s) pulsations, as well as the doublet seen in the $l = 1, k = -5$ (271.6 s) pulsation, can be used to infer the internal rotation profile of KIC08626021, or part of it. This is different from the usual approach (as used above) for estimating a rotation timescale only on the basis of the detected spacings within frequency multiplets. In that case, there is no information whatsoever on the location at which the rotation timescale originates. Instead, it is possible to probe the actual rotation profile as a function of depth^{20,30,45}.

The rotation kernels associated with each of these three frequency multiplets were computed on the basis of our optimal model and our mode identification. These functions play a similar role to the weight functions encountered above, except that they indicate the regions inside a model at which the local rotation law can be probed (or not). In the present case, we find that the outer 70% or so of the radius can be sounded for rotation. This corresponds to the outer roughly 81% of the mass of the star. Our results are summarized in Extended Data Fig. 6b. We find that the available multiplet data are compatible with the proposition that KIC08626021 rotates rigidly over its outer 70% or so in radius with a period of $P_{\text{rot}} = 46.34 \pm 2.54$ h. Given the available limited rotation data, we cannot exclude the possibility of some mild differential rotation about that value, but the most important result is that the star rotates globally (over at least 81% of its mass) very slowly by stellar standards. This adds to mounting evidence⁴⁶ that isolated white dwarfs have lost most of their initial angular momentum.

We can combine the value of the rotation period, P_{rot} , with our seismic estimate of the radius, R_* , of KIC08626021 to derive its equatorial velocity, V_{eq} , as follows: $V_{\text{eq}} = 2\pi R_*/P_{\text{rot}}$. Likewise, assuming that the star rotates rigidly over its full extent, from centre to surface, we can obtain a value for its total angular momentum, J , which is some 291 times smaller than the total angular momentum of the Sun (a future white dwarf). We report these data in Table 1. In terms of energy, these data correspond to a ratio of the total rotation energy of KIC08626021 to its internal energy of $E_{\text{rot}}/E_{\text{th}} = 8.11 \times 10^{-8}$, which is another way of demonstrating that rotation plays a negligible role in the further evolution and structure of this star.

Further observables. The potential of a hot V777 Her star such as KIC08626021 to serve as a benchmark for measuring plasmon neutrino emission is well known⁴⁷. Long-term efforts and stable pulsation frequencies are required in order to measure reliable values for the rates of period changes, dP/dt , associated with the secular evolution (cooling) of the star. Although the task appears difficult for KIC08626021 owing to its faintness ($K_p = 18.46$), we nevertheless decided to provide estimates of rates of period changes for the three highest-amplitude modes observed in that star. These may be useful in the future.

We thus computed two evolutionary sequences incorporating the detailed chemical profile inferred from our seismic analysis, one with neutrino losses included (as in a standard calculation) and the other without. The rates of period changes were computed directly from each of these sequences (Table 1). For each of the three retained modes (identified by their periods), we list the expected value of dP/dt with or without neutrino losses.

Excitation of the pulsation modes. It is obviously important to verify whether the pulsation modes of interest identified in the optimal model are indeed predicted to be excited through detailed stability calculations. For this, we use the Liège non-adiabatic code MAD⁴⁸, which takes properly into account the perturbations of the convective flux⁴⁹. As it is, with the optimal model obtained with the same convective efficiency of $ML2/\alpha = 1.25$ as used in the spectroscopic analyses of DB/DBA white dwarfs, we find no unstable modes. That is, the optimal model is located above the $ML2/\alpha = 1.25$ blue edge for dipole modes in the $T_{\text{eff}}\text{--}\log(g)$ diagram.

This situation is entirely analogous to the case of the cooler ZZ Ceti stars, for which it was found necessary to use a higher convective efficiency⁵⁰ in the pulsation models (the $ML2/\alpha = 1.0$ version) than that calibrated and used in the spectroscopic analyses ($ML2/\alpha = 0.6$) in order to best match the theoretical instability strip with the empirical (spectroscopic) one. The explanation is that the 'equivalent ML convective efficiency' increases with depth in white dwarfs, from the atmosphere down to the bottom of the convection zone, as demonstrated by detailed three-dimensional hydrodynamic calculations and recent calibration efforts⁵¹.

In the same way as for the ZZ Ceti stars⁵⁰, we examined the stability properties of models in all points similar to the optimal seismic model, but with an increased convective efficiency compared with the one used in the spectroscopic analyses.

Specifically, we considered $ML2/\alpha = 1.5$ and $ML2/\alpha = 1.6$. This is discussed above, but we recall here that the pulsation periods are quite insensitive to the choice of the convective efficiency in our models, while the driving and excitation of modes depend quite sensitively on the location of the base of the effective convection zone, a quantity that depends intimately on the choice of the convective efficiency.

Some of our results are presented in Extended Data Fig. 6c, which illustrates the ranges of detected periods in relation to the bands of predicted unstable modes. In particular, we find that a convective efficiency with the $ML2/\alpha = 1.6$ flavour is required to drive all of the eight observed modes in our model of KIC08626021. Additional calculations indicate that, for that particular efficiency (and for a total mass of $0.570M_{\odot}$), the blue edge of the V777 Her instability strip is located near $T_{\text{eff}} \approx 31,190$ K for the dipole modes, and around $T_{\text{eff}} \approx 32,090$ K for the quadrupole modes. This is probably sufficient also to account qualitatively for the presence of pulsations in the hottest known V777 Her pulsator, PG 0112+104 (ref. 43).

In short, and in retrospect, our extensive search for an optimal model in parameter space should have been done with a convective efficiency of, for example, $ML2/\alpha = 1.6$, to ensure a better consistency between adiabatic and non-adiabatic properties. However, for the reasons given above (mostly the insensitivity of the pulsation periods to the convective efficiency), we do not expect that the inferred chemical stratification would be much different to what we obtained should a new search be carried out. To be more precise, because this source of systematics has been considered in our evaluation of the error budget, any potential shift in the seismic solution is most likely within the conservative error estimates provided here.

Systematics. The procedure^{9,20} introduced to evaluate uncertainties represents, by definition, internal errors of the fit propagated to the derived parameters. Here, these internal errors are in fact very small, owing to the very high precision achieved in measuring the pulsation periods by the Kepler spacecraft and in matching the periods at this precision with a model. However, the internal chemical profile and global parameters inferred here do depend specifically on the particular components of the constitutive physics that went into our model building—an additional source of uncertainty that we call systematics or external errors. That is, the results must be sensitive to, for example, our choice of radiative opacity, conductive opacity, and equations of state for the fully ionized interior as well as the partially ionized non-ideal envelope. For example, had we used the Opacity Project (OP) radiative opacities instead of the OPAL data in our model codes, the inferred chemical profile would necessarily have been different. It is difficult to assess by how much, although we would expect only relatively small differences because the two sets of opacities have proven very similar, especially for the light elements such as hydrogen, helium, carbon and oxygen. To be more quantitative at this point would require a major computational effort (we would need to carry out many further full searches in parameter space, which is not feasible); such an effort will have to wait for the future, if deemed necessary. But other sources of systematics can be investigated in a more practical way and we discuss some of them below.

A more realistic core composition for a white dwarf should include a trace of neon-22, which is expected to be the most abundant of the minor species that must be present. Our approach to describing both the core and the envelope of a white dwarf has been limited to four elements—hydrogen, helium, carbon and oxygen—for which we have detailed, state-of-the-art tabular data covering the full density–temperature range of interest for the equation of state (liquid and solid phases), radiative opacity, and conductive opacity. These are the elements that leave a dominant seismic signature on the run of the Brunt–Väisälä (BV) frequency (because we are dealing with gravity modes), and, therefore, their abundances are the most easily measured through seismic analysis. In this context, we find that the presence of neon-22, in trace form, has no direct imprint on the BV frequency profile. Indeed, an examination of the neon-22 profile in figure 9 of ref. 52 indicates an essentially flat distribution, with a mass abundance of $X(^{22}\text{Ne}) = 0.02$ from the centre outwards to a $\log(q)$ of around -5 for their representative $0.5885M_{\odot}$ model of a white dwarf. Figure 5 of ref. 5 shows a similar result for their $0.64M_{\odot}$ model of a carbon–oxygen white dwarf calculated with the Modules for Experiments in Stellar Astrophysics (MESA) code. The profile of neon-22 being flat, its chemical signature on the BV frequency is limited only by the absolute value of that frequency, not by a composition ‘spike’ that would constitute a telltale sign of its presence. Given the small absolute abundance of neon-22, its presence will probably remain undetectable through seismic means.

Nonetheless, the (probable) presence of neon-22 must affect the pulsation periods, but the effect is certainly small *a priori*. It resides in the difference in constitutive physics between a dominant element, say oxygen, and the isotope neon-22. This difference can be seen as a correction to a correction, given the small abundance of that element to start with. We expect that neon-22’s largest effect is to cause a difference in the equation of state, through the effects of Coulomb interactions between the ions. The addition of a trace of neon-22 in a model ‘softens’ the equation of state; that is, for a given mass the radius will be slightly smaller.

This implies that the pulsation periods will systematically decrease by some small amount because the average density is slightly boosted.

The other effect comes from the conductive opacity (the radiative opacity of neon-22 is not important here, because neon-22 is present only in the degenerate core, where electron conduction dominates the energy-transfer process). Because of the larger average charge carried by neon-22 ions compared with, say, oxygen ions, we expect a slight increase in the global conductive opacity. This leads to a very small increase in the temperature, which reduces the pulsation periods of gravity modes in white dwarfs. This second phenomenon, a thermal effect, goes in the same direction—that is, a decrease in the pulsation periods—but it is less important than the first phenomenon, a mechanical effect. With some effort, we have been able to put these considerations on a firm quantitative basis. First, we computed an appropriate sophisticated equation of state for neon-22 that applies to the cores of white dwarfs, using a code developed at Université de Montréal⁵³. To our knowledge, this is the first detailed equation of state computed for an isotope such as neon-22. Next, we used the old conductive opacity code⁵⁴ to estimate the differences in conductive opacity between neon-22 and a reference element, which we take as oxygen. We note that we do not use the old Hubbard–Lampe conductive opacities in our model-building code. Instead, we use the latest upgrades⁵⁵ for hydrogen, helium, carbon and oxygen. Because a Potekhin version of the conductive opacity of neon-22 is not readily available, we resorted to the Hubbard–Lampe code to estimate in a differential sense the conductive opacity of neon-22 with respect to that of oxygen. We find that the conductive opacity of neon-22 is some 1.19 times larger than that of oxygen-16, averaged over the density–temperature domain of interest for our models of KIC08626021, that is, $6.8 < \log(T) < 7.9$, and $2.0 < \log(\rho) < 6.5$.

The next step was to build a standard reference model using a simple structure for a DB white dwarf: a pure oxygen core surrounded by a pure helium mantle. This was computed for a mass and an effective temperature equal to those inferred for KIC08626021. We pulsated that structure with our high-precision, finite-element adiabatic pulsation code and noted the pulsation periods. Next, we computed another similar DB equilibrium structure, except that the pure oxygen core was replaced by an oxygen-dominated core with a trace of neon ($X(^{22}\text{Ne}) = 0.02$) from the centre outwards to $\log(q) = -5$, following predicted evolutionary models⁵².

To do this, we ‘fooled’ our model-building code into ‘thinking’ that oxygen is carbon, while neon-22 is oxygen. That is, we replaced all of the constitutive physics appropriate for carbon by that of oxygen (equation of state and opacities), while we replaced the equation-of-state tabular data for oxygen with our new results for neon-22, and multiplied the modern (Potekhin) value of the conductive opacity of oxygen by the factor 1.19 described above.

As expected, there is no noticeable difference between the two models. The presence of neon-22 at the abundance used would not be revealed through seismology. However, as indicated above, we do find some tiny systematic differences in the absolute value of the BV frequency, which lead to slight differences in pulsation periods. Using the eight modes identified herein, we find that, on average, the pulsation periods are shorter by 0.34 s compared with the average period of 254.4 s, leading to a 0.1% effect—as expected, quite small. This comparison also confirms that most of the decrease in pulsation period is due to the difference in the equation of state; this is the main source of systematic errors. Given that a 0.1% shift in the periods of the optimal model translates into a S^2 value of about 0.12 for the period fit, we incorporated this value as a local minimal threshold of the merit function to compute the likelihood distribution, so the error estimates now reflect the impact of this external factor.

To test our specific choice of the convective efficiency $ML2/\alpha = 1.25$, we computed a model in all points similar to the optimal one found here, except that we adopted a reduced efficiency specified by the flavour $ML2/\alpha = 1.0$. The most important effect of this is to move the base of the outer helium convection zone from $\log(q) = -2.506$ to $\log(q) = -13.222$. Although this is important for non-adiabatic calculations because the driving engine is located at the base of the convection zone in a pulsating white dwarf of the V777 Her type, the pulsation periods are hardly affected given that the weight functions for the modes of interest have practically negligible amplitudes in these outermost layers. This expectation is explicitly verified by comparing the periods corresponding to the eight modes of interest in the two models. We find that the pulsation periods in the $ML2/\alpha = 1.0$ model are systematically and almost uniformly reduced by 0.0025% compared with those of the optimal ($ML2/\alpha = 1.25$) model. This is in line with the results depicted in figure 42 of ref. 17, which shows a small but nevertheless systematic increase in the pulsation period of a mode with increasing convective efficiency. We note that such small differences are still much larger than the precision of the Kepler data used here, meaning that a fully fledged search in parameter space would again be needed to fully measure the dependence of the inferred profile. This computed error budget is included in the estimate of our systematics.

Similar considerations can be made when addressing the issue that KIC08626021 is not really a pure DB white dwarf, but has an atmosphere that contains a small trace of hydrogen ($\log(\text{H}/\text{He}) = -3.0 \pm 0.1$), as indicated in our analysis of its optical spectrum. A more realistic model of KIC08626021 is that of a DBA star that contains that small trace of hydrogen in the outer convection zone. As above, we computed an extra model, similar in all points to our optimal model, but containing such a trace. This changes somewhat the structure of the convection zone, including a small shift of its base from $\log(q) = -12.506$ to $\log(q) = -12.533$. The effects are not negligible as far as the structure of the atmosphere itself is concerned. Indeed, because of the atmospheric opacity of hydrogen, there are differences between DB and DBA model atmospheres¹². But the consequences are again practically negligible as regards the pulsation periods because, as pointed out above, the weight functions have practically negligible amplitudes in the outer convection zone. Specifically, we find that the pulsation periods in the DBA model are systematically and almost uniformly reduced by 0.0010% compared with those of the optimal (DB) model. Incorporating these systematic errors as an external source in the computed error budget is done as before.

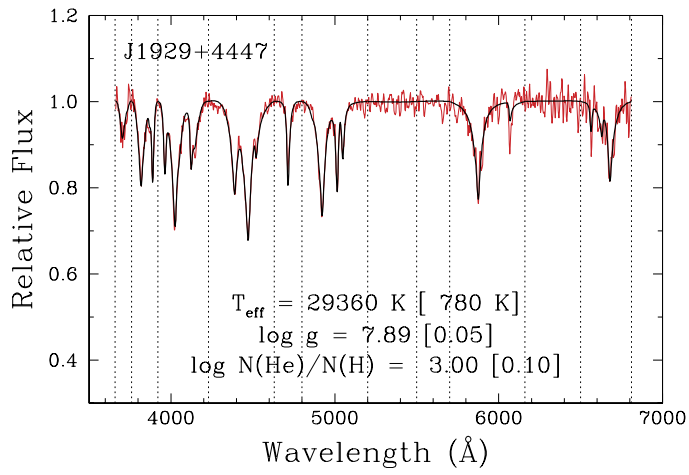
We also tested the impact of parameters that control how the adiabatic pulsation periods are computed, but we found this source of systematics to be much smaller than those discussed above. In the same vein, calculations of the non-adiabatic periods of the optimal model showed that non-adiabatic effects on the periods themselves are so small that they are irrelevant as a source of systematics in our results.

In summary, the error estimates that we provide with the derived parameters and chemical composition profiles include internal errors of the fit and a conservative estimate of the identified, most prominent sources of relevant external uncertainties. Of course, further systematics of yet unknown nature cannot be ruled out. The best means of identifying them will be to compare the properties of the seismic model with independent measurements of the stellar properties. We recall that the independent spectroscopic measurements of $\log(g)$ and T_{eff} do match well with the seismic solution, suggesting that other possible systematic errors are likely to be small. Further tests of solution accuracy should become possible when the full data from the GAIA global space astrometry project become available.

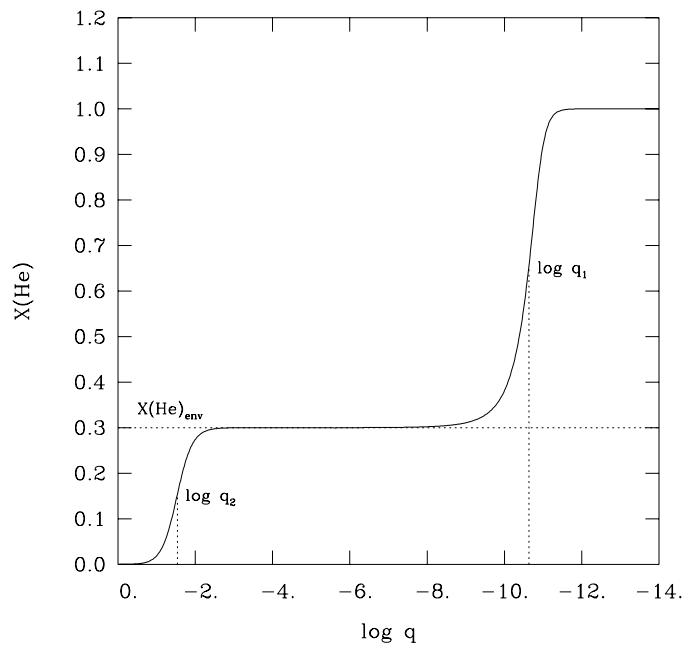
Code availability. We have opted not to make publicly available the highly specialized numerical codes used here because of their complexity.

Data availability. The data that support the findings of this study are available from the corresponding author upon request.

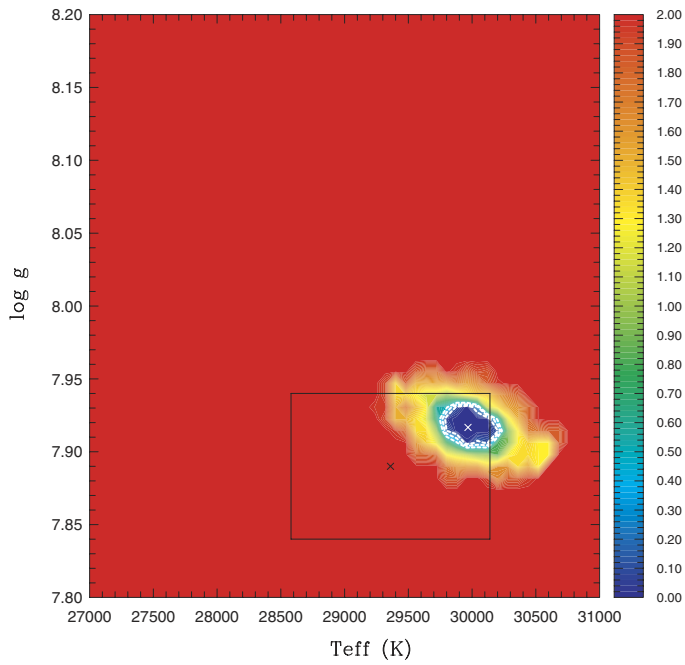
28. Brassard, P. *et al.* Discovery and asteroseismological analysis of the pulsating sdB star PG 0014+067. *Astrophys. J.* **563**, 1013–1030 (2001).
29. Charpinet, S., Fontaine, G., Brassard, P., Green, E. M. & Chayer, P. Structural parameters of the hot pulsating B subdwarf PG 1219+534 from asteroseismology. *Astron. Astrophys.* **437**, 575–597 (2005).
30. Charpinet, S. *et al.* Testing the forward modeling approach in asteroseismology. II. Structure and internal dynamics of the hot B subdwarf component in the close eclipsing binary system PG 1336-018. *Astron. Astrophys.* **489**, 377–394 (2008).
31. Van Grootel, V., Charpinet, S., Brassard, P., Fontaine, G. & Green, E. M. Third generation stellar models for asteroseismology of hot B subdwarf stars. A test of accuracy with the pulsating eclipsing binary PG 1336-018. *Astron. Astrophys.* **553**, A97 (2013).
32. Charpinet, S., Giammichele, N., Brassard, P., Van Grootel, V. & Fontaine, G. Method and tools for an objective approach of white dwarf asteroseismology. *Astron. Soc. Pac. Conf. Ser.* **493**, 151–155 (2015).
33. Beauchamp, A. *et al.* Spectroscopic studies of DB white dwarfs: the instability strip of the pulsating DB (V777 Herculis) stars. *Astrophys. J.* **516**, 887–891 (1999).
34. Brassard, P., Pelletier, C., Fontaine, G. & Wesemael, F. Adiabatic properties of pulsating DA white dwarfs. III—a finite-element code for solving nonradial pulsation equations. *Astrophys. J. Suppl. Ser.* **80**, 725–752 (1992).
35. Brassard, P. & Charpinet, S. PULSE: a finite element code for solving adiabatic nonradial pulsation equations. *Astrophys. Space Sci.* **316**, 107–112 (2008).
36. Østensen, R. H. Nine months of monitoring of a V777-Her pulsator with the Kepler spacecraft. *Astron. Soc. Pac. Conf. Ser.* **469**, 3–7 (2013).
37. Bischoff-Kim, A. & Østensen, R. H. Asteroseismology of the Kepler field DBV white dwarf. It is a hot one. *Astrophys. J.* **742**, L16 (2011).
38. Corsico, A. H., Althaus, L. G., Miller Bertolami, M. M. & Bischoff-Kim, A. Asteroseismology of the Kepler V777 Herculis variable white dwarf with fully evolutionary models. *Astron. Astrophys.* **541**, A42–A50 (2012).
39. Brassard, P. & Fontaine, G. The case of the Kepler DBAV star J1929+4447. *EPJ Web Conf.* **43**, 05010 (2013).
40. Koester, D. White dwarf spectra and atmosphere models. *Mem. Soc. Astron. Ital.* **81**, 921 (2010).
41. Bergeron, P., Saffer, R. & Liebert, J. A spectroscopic determination of the mass distribution of DA white dwarfs. *Astrophys. J.* **394**, 228–247 (1992).
42. Van Grootel, V., Fontaine, G., Brassard, P. & Dupret, M.-A. The theoretical instability strip of V777 Her white dwarfs. *Astron. Soc. Pac. Conf. Ser.* **509**, 321 (2017).
43. Hermes, J. J. *et al.* A deep test of radial differential rotation in a helium-atmosphere white dwarf. I. Discovery of pulsations in PG 0112+104. *Astrophys. J.* **835**, 277 (2017).
44. Koester, D., Provencal, J. & Gänsicke, B. Atmospheric parameters and carbon abundance for hot DB white dwarfs. *Astron. Astrophys.* **568**, A118 (2014).
45. Charpinet, S., Fontaine, G. & Brassard, P. Seismic evidence for the loss of stellar angular momentum before the white-dwarf stage. *Nature* **461**, 501–503 (2009).
46. Fontaine, G., Brassard, P. & Charpinet, S. The angular momentum of isolated white dwarf stars. *Astron. Soc. Pacific* **469**, 115 (2013).
47. Sullivan, D. J. Time-series spectroscopy and photometry of the helium atmosphere pulsating white dwarf EC 20058-5234. *Astron. Soc. Pac. Conf. Ser.* **509**, 315 (2017).
48. Dupret, M.-A. Nonradial nonadiabatic stellar pulsations: a numerical method and its application to a beta Cephei model. *Astron. Astrophys.* **366**, 166–173 (2001).
49. Van Grootel, V. *et al.* The instability strip of ZZ Ceti white dwarfs. I. Introduction of time-dependent convection. *Astron. Astrophys.* **539**, A87–A96 (2012).
50. Van Grootel, V., Fontaine, G., Brassard, P. & Dupret, M.-A. The newly discovered pulsating low-mass white dwarfs: an extension of the ZZ Ceti instability strip. *Astrophys. J.* **762**, 57 (2013).
51. Tremblay, P.-E. *et al.* Calibration of the mixing-length theory for convective white dwarf envelopes. *Astrophys. J.* **799**, 142 (2015).
52. Althaus, L. G. *et al.* The formation and evolution of hydrogen-deficient post-AGB white dwarfs: the emerging chemical profile and the expectations for the PG 1159-DB-DQ evolutionary connection. *Astron. Astrophys.* **435**, 631–648 (2005).
53. Kitsikis, A., Fontaine, G. & Brassard, P. Determination of a modern equation of state for the liquid/solid core of white dwarf stars. *Astron. Soc. Pac. Conf. Ser.* **334**, 65 (2005).
54. Hubbard, W.-B. & Lampe, M. Thermal conduction by electrons in stellar matter. *Astrophys. J. Suppl. Ser.* **18**, 297 (1969).
55. Cassisi, S., Potekhin, A. Y., Pietrinferni, A., Catelan, M. & Salaris, M. Updated electron-conduction opacities: the impact on low-mass stellar models. *Astrophys. J.* **661**, 1094–1104 (2007).



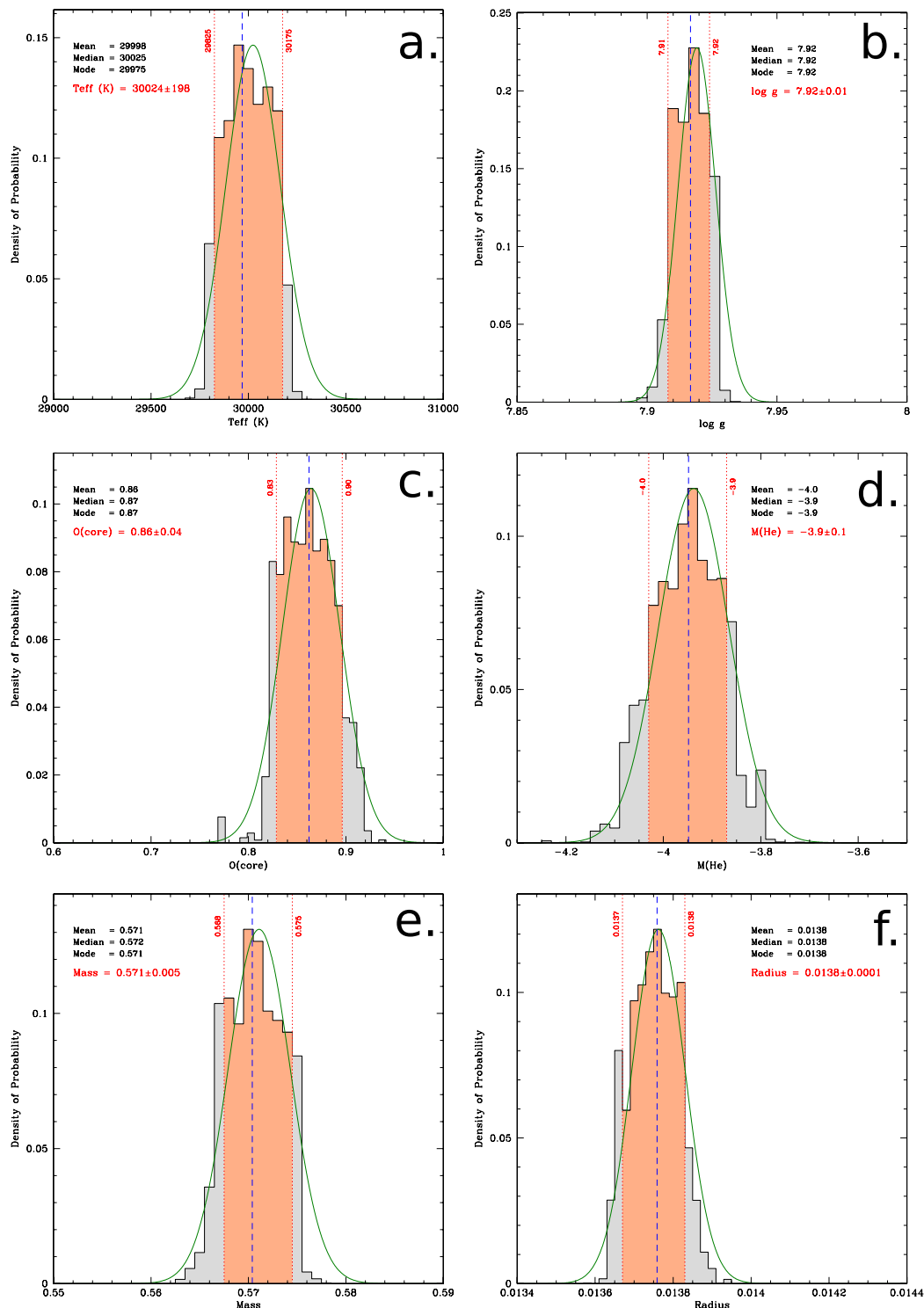
Extended Data Figure 1 | Model fit to the spectrum of KIC08626021 (catalogued as J1929+4447). The model fit is shown in black, the observed spectrum in red. Our estimates indicate that KIC08626021 is the second hottest of the known pulsators of the V777 Her type. The quoted uncertainties are only the formal errors of the fit.



Extended Data Figure 2 | Parameterization of the helium profile in the envelope of a typical DB white dwarf model. The figure shows the local helium mass fraction, $X(\text{He})$, as a function of the fractional mass depth, $\log(q) \equiv \log[1 - M(r)/M]$. Along with the three quantities depicted in the plot— $\log(q_1)$, $\log(q_2)$ and $X(\text{He})_{\text{env}}$ as defined previously—there are two other hidden parameters that are related to the shape of the helium profile in the descent centred on $\log(q_1)$ and in the descent centred on $\log(q_2)$ (see text for details).

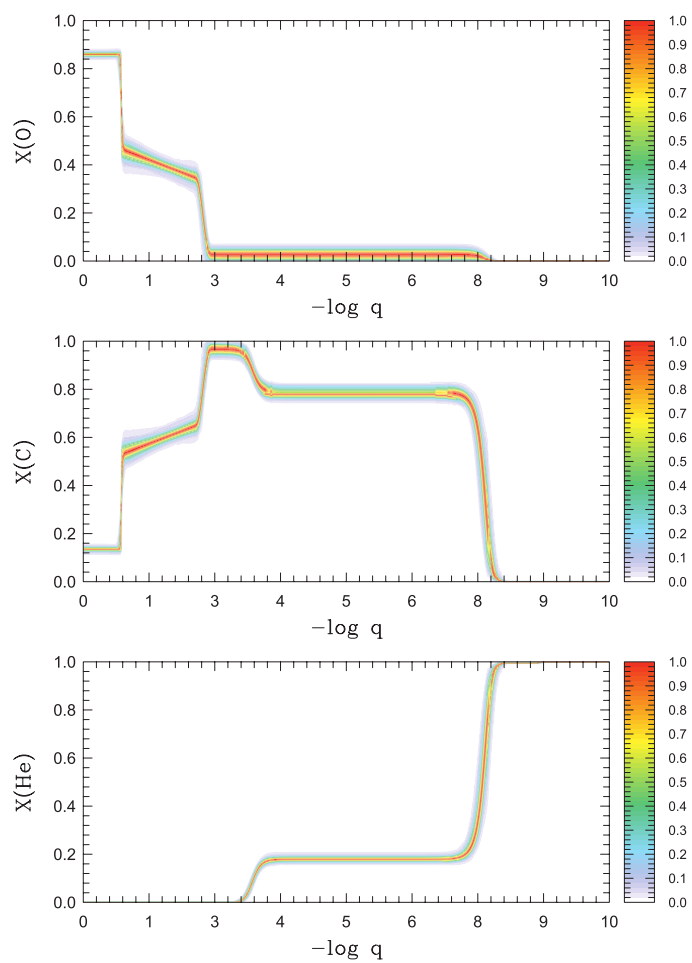


Extended Data Figure 3 | Map of the 15-dimensional merit function S^2 projected onto the $\log(g)$ – T_{eff} plane for models of KIC08626021. The merit function is shown on a logarithmic colour scale (base 10). The location of the optimal model in this plane is indicated by a white cross. The white dotted curves delimit the regions where the merit function has values within the 1σ , 2σ and 3σ confidence levels relative to the best-fitting solution. The black cross surrounded by the solid black box indicates the independent spectroscopic solution and its 1σ uncertainties.

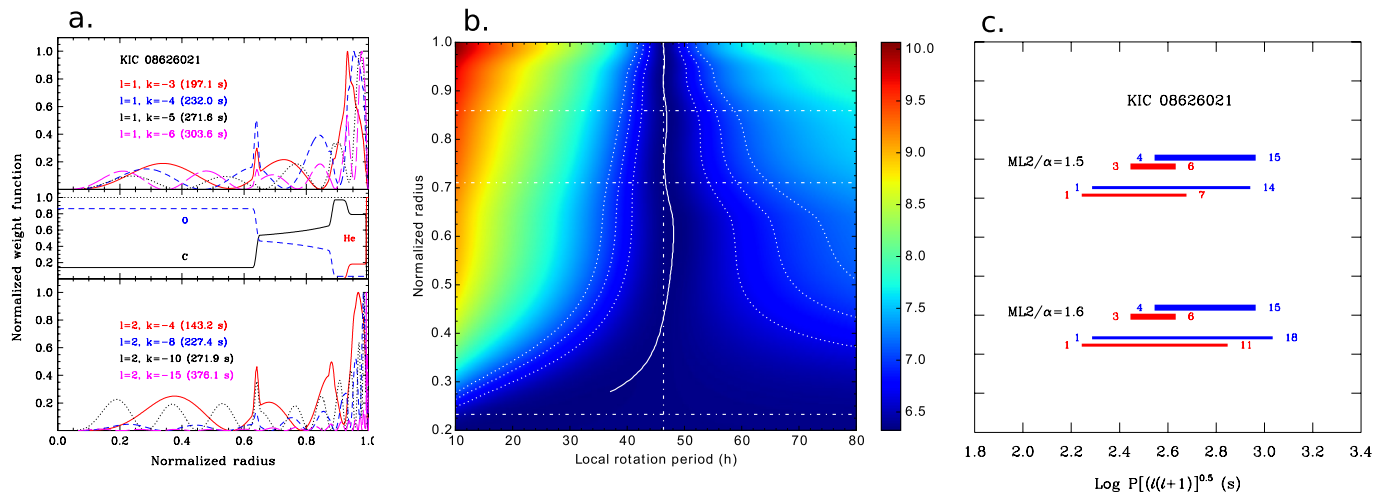


Extended Data Figure 4 | Results of the statistical analysis carried out in parameter space about the optimal seismic model for KIC08626021. Only the most interesting parameters are illustrated here. Each histogram shows the derived probability-density function for a given model parameter. The orange hatched region between the two vertical solid red lines defines the $\pm 1\sigma$ range, containing 68.3% of the distribution. The green curve defines the Gaussian fit applied to the distribution, which gives the $\pm 1\sigma$ range. The blue vertical dashed line indicates the value of the

parameter of the optimal model solution. The mean, median and mode values are also indicated. The estimate of the parameter of interest, indicated in red, is the statistical value (not the optimal one) and corresponds to the central value of the $\pm 1\sigma$ interval. The various panels correspond to the following model parameters: **a**, effective temperature; **b**, surface gravity; **c**, central mass fraction of oxygen; **d**, total mass fraction of helium; **e**, total mass; and **f**, total radius of the star (in logarithmic units).



Extended Data Figure 5 | Derived probability-distribution functions, normalized to 1, for the oxygen, carbon and helium profiles.



Extended Data Figure 6 | Various properties of our optimal model.

a. Normalized weight function plotted against the normalized radius. Each individual weight function for the eight gravity modes of interest from the optimal model of KIC08626021 is normalized to a maximum value of 1.0. The weight function of a mode indicates the layers contributing most to the integral, giving the frequency (period) of the mode according to a well known variational principle in linear pulsation theory. The middle panel illustrates the chemical stratification of the model. All of the identified modes are useful probes of the core composition. **b.** Internal rotation profile. Contour map of the two-dimensional merit function that optimizes the match between the observed spacings in the three frequency multiplets with computed spacings on the basis of our seismic model. This is shown in terms of depth (expressed as the normalized radius) and in terms of the local rotation period of the inner region in the two-zone approach of ref. 45. The best-fitting solution is illustrated by the nearly vertical white curve about the solid-body solution (vertical dot-dashed

white line). The dotted white curves on both sides of the solution depict its associated 1σ , 2σ and 3σ uncertainty contours. The fact that these contours diverge out at the greater depths considered here indicates that the rotationally split gravity modes available here lose their capacity to measure the local rotation rate at greater depths (indeed, their rotation kernels have negligible amplitudes at such depths). The horizontal white dot-dashed lines indicate the layer below which there is 99%, 90% and 10% of the mass of the star, from top to bottom. **c.** Comparison between the ranges of detected and excited periods in KIC08626021. The panel shows the detected periods (thick lines) with the bands of excited periods (thin lines) in two models similar to the optimal seismic model, but computed with the higher convective efficiency of $ML2/\alpha = 1.5$ (above) and $ML2/\alpha = 1.6$ (below). The reduced period is used as the abscissa in order to have comparable values for both dipole mode (in red) and quadrupole mode (in blue). The radial order k is indicated at both ends of each range.

Extended Data Table 1 | Mode identification and details of the frequency fit obtained for the optimal solution

l	k	ν_{obs} (μHz)	ν_{th} (μHz)	P_{obs} (s)	P_{th} (s)	$\log E$ (erg)	C_{kl}	ID
1	−1	...	8068.03208	...	123.945962	48.875	0.2874	
1	−2	...	6067.64155	...	164.808681	47.711	0.4012	
1	−3	5073.23411	5073.23411	197.112922	197.112922	47.181	0.4276	f_2
1	−4	4309.91490	4309.91490	232.023143	232.023143	46.678	0.4555	f_1
1	−5	3681.80286	3681.80286	271.606068	271.606068	46.197	0.4681	f_3
1	−6	3294.36928	3294.36928	303.548241	303.548241	46.018	0.4702	f_6
1	−7	...	2971.71850	...	336.505628	45.878	0.4784	
1	−8	...	2616.20375	...	382.233227	45.571	0.4834	
2	−2	...	9825.05526	...	101.780598	47.427	0.0963	
2	−3	...	8088.37817	...	123.634180	46.913	0.1098	
2	−4	6981.26129	6981.26129	143.240592	143.240592	46.557	0.1218	f_9
2	−5	...	6110.75661	...	163.645857	46.055	0.1411	
2	−6	...	5399.55096	...	185.200586	45.886	0.1440	
2	−7	...	4859.14357	...	205.797583	45.930	0.1417	
2	−8	4398.37230	4398.37230	227.356834	227.356834	45.439	0.1540	f_5
2	−9	...	3994.39516	...	250.350794	45.414	0.1528	
2	−10	3677.99373	3677.99373	271.887358	271.887358	45.365	0.1520	f_7
2	−11	...	3406.57028	...	293.550380	44.746	0.1593	
2	−12	...	3138.95972	...	318.576882	44.555	0.1575	
2	−13	...	2967.78647	...	336.951465	44.208	0.1581	
2	−14	...	2808.26351	...	356.091939	43.571	0.1622	
2	−15	2658.77740	2658.77740	376.112721	376.112721	43.540	0.1619	f_4
2	−16	...	2531.88338	...	394.962900	43.834	0.1597	

$\log E$, kinetic energy of the mode (in logarithmic scale); C_{kl} , first-order solid-body rotation coefficient.

Extended Data Table 2 | Defining parameters of the optimal seismic model found for KIC08626021

Quantity	Optimum value
T_{eff} (K)	29,968
$\log g$	7.9167
$\log q_1$	-7.630
$\log q_2$	-3.229
$X(\text{He})_{\text{env}}$	0.181
P_1	6.664
P_2	13.246
Core O	0.862
t_1	-0.728
Δt_1^*	0.0294
$t_1(\text{O})$	0.4655
t_2	-2.284
Δt_2	0.132
$t_2(\text{O})$	0.349
$X_{\text{env}}(\text{O})$	0.0294

*A secondary parameter is defined by the relation $\log(q_3) = t_1 + \Delta t_1/2$, representing the extent in $\log(q)$ of the inner flat portion of the oxygen profile in Fig. 2. The parameter $\log(q_3)$ is used in the comparison with the results of other investigations.

Extended Data Table 3 | Observed periods, mode identification and normalized merit functions for past seismic studies of KIC08626021 compared with this work

Mode	143.24 (s)	197.11 (s)	227.36 (s)	232.02 (s)	271.61 (s)	271.89 (s)	303.55 (s)	376.11 (s)	s^2 (s ²)	References
$l, -k$...	1,3	...	1,4	1,5	...	1,6	2,15	7.84×10^{-2}	37
$l, -k$...	1,3	...	1,4	1,5	...	1,6	1,8	3.74	38
$l, -k$...	1,3	...	1,4	1,5	...	1,6	1,8	2.93×10^{-2}	39*
$l, -k$...	1,3	...	1,4	1,5	...	1,6	2,15	2.51×10^{-5}	39†
$l, -k$	2,4	1,3	2,8	1,4	1,5	...	1,6	1,8	1.31×10^{-1}	21‡
$l, -k$	2,4	1,3	2,8	1,4	1,5	...	1,6	1,8	4.80×10^{-1}	21§
$l, -k$	2,4	1,3	2,8	1,4	1,5	2,10	1,6	2,15	1.75×10^{-16}	This work

Previous results from refs 21 and 37–39.

*Solution 1 of ref. 39.

†Solution 2 of ref. 39.

‡Solution 1 of ref. 21.

§Solution 2 of ref. 21.

Extended Data Table 4 | Main inferred parameters from past seismic studies of KIC08626021 in comparison with this work

$\log g$ (cm s^{-2})	T_{eff} (K)	M_*/M_{\odot}	$\log M(\text{He})/M_*$	$\log q_1$	$\log q_2$	$\log q_3$	$X(\text{O})_{\text{center}}$	References
...	29200	0.570	...	−6.30	−2.80	−0.19	0.60–0.65	37
8.099	27263	0.664	−2.27	−5.95	−1.63	−0.27	0.65	38
7.840	29990	0.539	−4.29	−7.59	−3.59	...	0.09	39*
7.900	28750	0.563	−1.91	−6.22	−0.91	...	0.89	39†
...	29650	0.550	...	−7.90	−3.10	−0.11	0.55	21‡
...	29350	0.550	...	−8.40	−3.00	−0.11	0.70	21§
7.917(9)	29968(150)	0.570(4)	−3.95(3)	−7.63(9)	−3.23(5)	−0.72(1)	0.86(2)	This work

Previous results from refs 21 and 37–39.

*Solution 1 of ref. 39.

†Solution 2 of ref. 39.

‡Solution 1 of ref. 21.

§Solution 2 of ref. 21.

Centimetre-scale electron diffusion in photoactive organic heterostructures

Quinn Burlingame^{1*}, Caleb Coburn^{2*}, Xiaozhou Che³, Anurag Panda⁴, Yue Qu¹ & Stephen R. Forrest^{1,2,3,4}

The unique properties of organic semiconductors, such as flexibility and lightness, are increasingly important for information displays, lighting and energy generation. But organics suffer from both static and dynamic disorder, and this can lead to variable-range carrier hopping^{1,2}, which results in notoriously poor electrical properties, with low electron and hole mobilities and correspondingly short charge-diffusion lengths of less than a micrometre^{3,4}. Here we demonstrate a photoactive (light-responsive) organic heterostructure comprising a thin fullerene channel sandwiched between an electron-blocking layer and a blended donor:C₇₀ fullerene heterojunction that generates charges by dissociating excitons. Centimetre-scale diffusion of electrons is observed in the fullerene channel, and this can be fitted with a simple electron diffusion model. Our experiments enable the direct measurement of charge diffusivity in organic semiconductors, which is as high as 0.83 ± 0.07 square centimetres per second in a C₆₀ channel at room temperature. The high diffusivity of the fullerene combined with the extraordinarily long charge-recombination time yields diffusion lengths of more than 3.5 centimetres, orders of magnitude larger than expected for an organic system.

A series of devices was fabricated with the structure shown in Fig. 1a and described in Methods, with a device photograph shown in Fig. 1b (inset). Devices are identified following the convention: (donor in the heterojunction)-(type and thickness of the neat fullerene channel)-(electron-blocking layer: 'neat' or 'mixed'). Neat electron-blocking layers comprise only 8 nm of bathophenanthroline (BPhen), whereas the mixed electron blockers consist of 10 nm of 1:1 volume ratio BPhen:C₆₀, with 5 nm of BPhen on top. The devices share a common architecture with organic photovoltaic cells (OPVs) with a planar-mixed donor/acceptor (D/A) heterojunction^{5,6}, whose power conversion efficiencies are over 9% (ref. 7). Using the experimental set-up and device illustrated in Fig. 1a, we measured the steady-state photocurrent response of an OPV comprising the donor molecule^{7,8} 2-((7-(5-(dip-tolylamino)thiophen-2-yl)benzo[c][1,2,5]thiadiazol-4-yl)methylene) malononitrile (DTDCPB) blended with C₆₀, as a function of the excitation position under approximately 0.15 mW cm^{-2} continuous illumination at a wavelength of 633 nm through the substrate from a fibre-coupled He-Ne laser. DTDCPB has a donor-acceptor-acceptor' (d-a-a') structure in which an electron-donating moiety and an electron-withdrawing moiety are bridged by another electron-accepting block. Results are shown in Fig. 1b.

In contrast to typical OPVs that exhibit no photoresponse to light incident outside the area of overlap of their anode and cathode, the DTDCPB-(10 nm C₆₀)-neat device generated a substantial photoresponse to such illumination. The magnitude of the steady-state photocurrent just outside the cathode was 40% of the peak within the device, decreasing approximately linearly to 12% at a distance $L = 10 \text{ mm}$ away from the cathode edge, probably owing to recombination or trapping at the film edge where it contacted the encapsulation epoxy. The large

drop in signal intensity at the edge of the cathode originates from a decrease in light absorption due to the lack of cathode reflection as well as a decrease in efficiency of charge generation and collection due to the lack of a built-in field outside the contact area. The photocurrent generated outside the contact overlap, henceforth called 'channel current', can cause overestimation of the short-circuit current in an OPV when it is overfilled by the illumination source⁹. Devices with 2-((7-(4-[N,N-bis(4-methylphenyl)amino]phenyl)-2,1,3-benzothiadiazol-4-yl)methylene) propane-dinitrile (DTDCPB) as the donor also exhibited channel currents, whereas those with boron subphthalocyanine chloride (SubPc) or tetraphenylidibenzoperiflanthene (DBP) as the donor had no response to light outside the cathode edge.

The transient behaviour of channel currents in all device architectures was investigated using the experimental set-up in Fig. 1a, with 500 μs pulses of 405-nm-wavelength light as the illumination source. Figure 1c shows the transient channel current from the DTDCPB-(10 nm C₆₀)-neat device illuminated at $L = 1-10 \text{ mm}$. As L increased, the amplitude and arrival time of the channel current varied

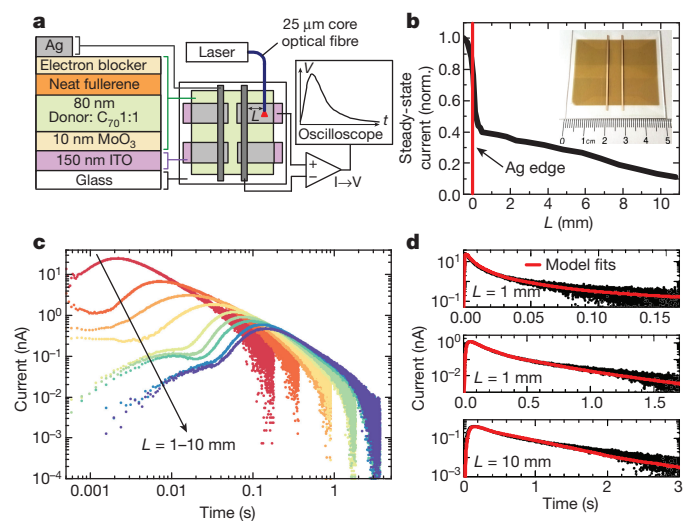


Figure 1 | Device structure, experimental set-up and distance-dependent photocurrent. **a**, Device structure (far left, not to scale) and schematic of the substrate design with experimental set-up for transient photocurrent measurements (right), where L is the distance between the fibre illumination position and the edge of the silver cathode. **b**, Normalized steady-state current of a DTDCPB-(10 nm C₆₀)-neat device due to constant 633 nm illumination at position L . Inset, photograph of the device before encapsulation. **c**, Room-temperature photocurrent transients of a DTDCPB-(10 nm C₆₀)-neat device illuminated with 500 μs pulses at a wavelength of 405 nm for $L = 1-10 \text{ mm}$. **d**, Model simulations of charge diffusion (lines) and corresponding transient photocurrent data (points) at $L = 1 \text{ mm}$, 5 mm and 10 mm .

¹Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, USA. ²Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA. ³Applied Physics Program, University of Michigan, Ann Arbor, Michigan 48109, USA. ⁴Department of Materials Science and Engineering, University of Michigan, Ann Arbor, Michigan 48109, USA.

*These authors contributed equally to this work.

Table 1 | Room-temperature charge-diffusion parameters extracted from distance-dependent transient current measurements

Device	D (cm ² s ⁻¹)	k (s ⁻¹)	μ^* (cm ² V ⁻¹ s ⁻¹)	Q (nC)
DTDCPB-(10 nm C ₆₀)-neat	0.83 ± 0.07	0.9 ± 0.6	32 ± 3	2.2 ± 0.8
DTDCTB-(10 nm C ₆₀)-neat	0.67 ± 0.06	1 ± 1	26 ± 2	0.38 ± 0.02
DTDCTB-(5 nm C ₆₀)-neat	0.53 ± 0.03	0.4 ± 0.4	20 ± 1	0.35 ± 0.02
DTDCTB-(2 nm C ₆₀)-neat	0.16 ± 0.02	0.7 ± 0.7	6 ± 1	0.29 ± 0.03
DTDCTB-(10 nm C ₆₀)-mixed	0.37 ± 0.08	0.3 ± 0.2	14 ± 3	3.2 ± 0.8
DTDCTB-(10 nm C ₇₀)-neat	0.16 ± 0.01	2.4 ± 0.8	6 ± 1	0.21 ± 0.02

*Estimated using the Einstein relation.

by nearly two orders of magnitude, although the standard deviation of the integrated charge collected for each transient was less than 10% of the mean for all transients. That is, we observe no measurable loss in the total number of charges collected, independent of the position of excitation, indicating that the charge-diffusion length L_D in the structure is considerably greater than the device length of 1 cm. In devices with C₇₀ channels, the integrated signal decreased by 50% over 5 mm, suggesting that L_D is small compared with that for C₆₀. The external quantum efficiency (EQE, the number of electrons collected per incident photon) decreased as a function of pump pulse energy and duration, presumably owing to increased recombination at higher polaron concentrations. In the DTDCTB-(10 nm C₆₀)-neat device at $L = 2$ mm, the EQE decreased from 30% to 15% as the pump pulse energy increased from 0.11 nJ to 1.7 nJ at a wavelength of 637 nm, and EQE decreased by 72% as the pulse length was increased from 0.1 ms to 100 ms. The collection efficiency of channel currents was wavelength-independent: that is, it tracked the absorption spectrum of the blended heterojunction. Channel currents were observed only when illuminating the organic area above the indium tin oxide (ITO) anode, which was needed to collect the photogenerated holes, thus preventing sample charging and reducing recombination.

A simple charge-diffusion model described in Methods was used to fit all transient currents as a function of L . Detailed transients (points) from Fig. 1c are plotted on a logarithmic-linear scale for the DTDCTB-(10 nm C₆₀)-neat device in Fig. 1d at $L = 1$ mm, 5 mm and 10 mm. The parameters extracted from the fits (lines) are given in Table 1. Among devices grown in the same batch, DTDCTB and DTDCPB devices with 10-nm-thick C₆₀ channel and neat electron-blocking layers had comparable diffusivities, which were reduced by replacing the C₆₀ with C₇₀, and by replacing the neat electron-blocking layer with a mixed layer. This reduction is presumably due to electron diffusion into the mixed layer where the diffusivity D is relatively low. Decreasing the C₆₀ channel thickness in the DTDCTB-C₆₀-neat devices from 10 nm to 2 nm also decreased D and the amount of total charge, Q , injected into the fullerene layer. The peak-to-peak roughness of the films grown on ITO is typically several nanometres as measured by atomic force microscopy¹⁰; thus, the thin fullerene channels are likely to have discontinuities and thickness variations that disrupt electron diffusion in thinner channels. The upper bound for the rate of charge trapping and recombination, k , in C₇₀ was about five times as high as that in C₆₀ channels. There were batch-to-batch variations in D , k and Q , but the relative performance between architectures was consistent; that is, devices with neat blockers or neat C₆₀ always had considerably higher D than devices with mixed blockers or C₇₀ grown in the same batch.

Ultraviolet photoelectron spectra were measured for donor:C₇₀ blends both with and without a cap of 5 nm C₆₀; see Methods. With the DTDCPB donor, we observe a difference of 0.42 ± 0.1 eV between the highest occupied molecular orbital (HOMO) energy (E_{HOMO}) of the blended C₇₀ and neat C₆₀ cap, whereas the difference in E_{HOMO} between neat C₆₀ and C₇₀ is about 0.1 eV. The difference in E_{HOMO}

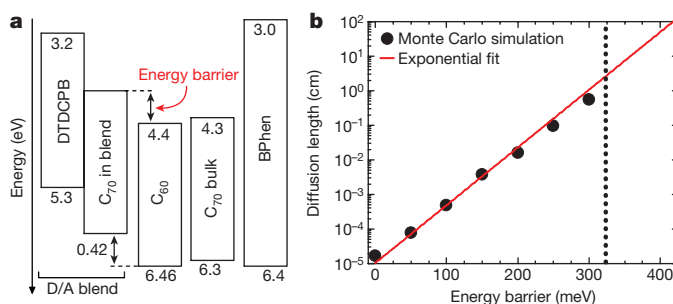


Figure 2 | Energetics of materials used in devices. **a**, Energy-level diagram extracted from ultraviolet photoelectron spectroscopy measurements on a 10-nm-thick C₇₀ and a 10-nm-thick DTDCPB:C₇₀ blended film with and without a 5-nm-thick C₆₀ film grown on its surface. **b**, Monte Carlo simulations of the hopping diffusion length as a function of the energy barrier height for electrons between the neat channel layer and the donor-acceptor heterojunction. The energy barrier is defined in **a**. The dotted line indicates the lower bound of the energy barrier measured by ultraviolet photoelectron spectroscopy.

between DBP:C₇₀ blends with and without a 5 nm C₆₀ cap was less than 100 meV, indicating that DBP does not noticeably shift E_{HOMO} of C₇₀ when the two constituents are blended.

To quantify the impact of the energy barrier between the fullerene channel and heterojunction, we performed Monte Carlo simulations of the carrier transport within the device structure to calculate L_D (see Methods), with results shown in Fig. 2b. We find that L_D is thermally activated as a function of the barrier height, E_B , following L_D (μm) = $(0.1 \pm 0.03)\exp(E_B/k_B T)$, where k_B is Boltzmann's constant and T is the temperature. The fit is indicated by the solid line in Fig. 2b. The pre-factor is the diffusion length at $E_B = 0$. The lower bound of L_D is 2.2 cm at 320 meV (dotted line, Fig. 2b), which corresponds to the lower bound of the measured energy barrier in the DTDCPB:C₇₀ film with a C₆₀ cap. The L_D inferred from the measured $E_B = 0.42$ eV is two orders of magnitude larger than this lower bound.

The temperature dependence of the current transients in DTDCTB-(10 nm C₆₀)-neat and DTDCTB-(10 nm C₇₀)-neat devices was obtained at 20 K intervals in the range $300 \text{ K} \geq T \geq 120 \text{ K}$. The results are shown in Fig. 3a and b at $L = 2$ mm and $L = 1$ mm, respectively. Each current transient was fitted using the charge-diffusion model as shown by solid lines in Fig. 3c and d, with the extracted values of D and k plotted against $1,000/T$ in Fig. 3e and f.

The lifetime of electrons in the channel is determined by the rates of trapping and recombination at defects, and by thermionic emission into the heterojunction where recombination can occur. A sufficiently high heterojunction energy barrier and low defect densities are therefore required to enable transport over macroscopic distances. In devices with d-a-a' donors, the energy levels of C₇₀ undergo a polarization shift due to the high dipole moments of DTDCTB and DTDCPB (14.5 debye and 12.0 debye, respectively)^{7,8}. This shift forms the required energy barrier at the C₆₀ channel/heterojunction interface (0.42 ± 0.1 eV for DTDCPB, as shown in Fig. 2a) that confines electrons within the channel. Monte Carlo simulations confirm that this barrier supports centimetre-scale diffusion, whereas DBP or SubPc donor devices with barrier heights of less than 100 meV have a much smaller $L_D < 5$ μm. The centimetre-scale L_D observed in devices with d-a-a' donors suggests that, in addition to the large E_B , the channel and its interfaces have a remarkably low density of deep electron traps and recombination centres. This is surprising for fullerenes, which, despite their unusually high mobility and diffusivity among molecular solids^{11–13}, form disordered and phase-separated amorphous and crystalline domains¹⁴.

Long-range electron diffusion was also observed to circumvent barriers introduced by physically cutting through the channel (Extended Data Fig. 1), as well as in an electron-only charge-injecting sample (Extended Data Fig. 2). Indeed, measurements of D and k in these experiments are completely consistent with values obtained through

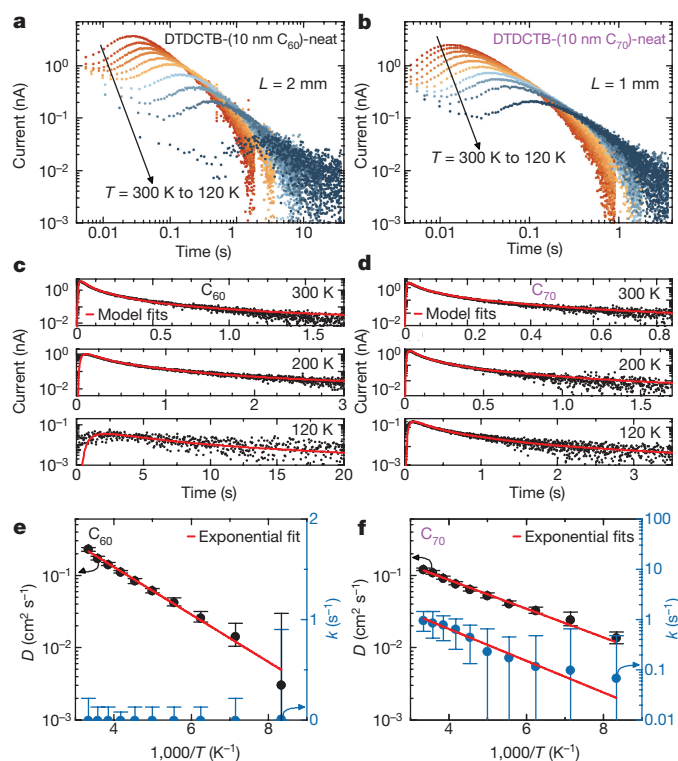


Figure 3 | Temperature dependence of channel currents.

a, b, Temperature-dependent transient photocurrent data at 20 K intervals from 300 K to 120 K in response to 2-ms-long pulses of 637-nm-wavelength illumination on a DTDCTB-(10 nm C_{60})-neat device at $L = 2$ mm (**a**) and a DTDCTB-(10 nm C_{70})-neat device at $L = 1$ mm (**b**). **c, d**, Data (points) and corresponding charge-diffusion model fits (lines) at 300 K, 200 K and 120 K for the DTDCTB-(10 nm C_{60})-neat (**c**) and the DTDCTB-(10 nm C_{70})-neat (**d**) devices. **e, f**, Diffusivity D (black points) and k (blue points) versus temperature T , extracted from the simulations in **c** and **d** for the DTDCTB-(10 nm C_{60})-neat (**e**) and the DTDCTB-(10 nm C_{70})-neat (**f**) devices. Lines show fits to the parameters.

photogeneration in Figs 1 and 2. We note that drift-dominated lateral spreading of charges over long time periods has been observed in unipolar devices at organic/insulator interfaces owing to the lack of recombination¹⁵. The results presented here are inherently different because the electron transport in ref. 15 is entirely diffusive, is associated with the bulk and is observed despite the presence of optically generated holes.

Replacing the 10 nm layer of C_{60} with C_{70} reduces D from 0.67 ± 0.06 $\text{cm}^2 \text{s}^{-1}$ to 0.16 ± 0.01 $\text{cm}^2 \text{s}^{-1}$ at room temperature. The Einstein relation, $D = \mu k_B T / q$, where q is the electron charge and μ is mobility, suggests that the larger D observed in the C_{60} devices is consistent with its higher mobility^{16,17}. This relation can also be used to estimate the electron mobility of each device, as listed in Table 1, although we note that D/μ in some organic systems has been shown to be larger than that predicted by the Einstein relation, because of disorder in the conduction-site energies¹⁸. We find that the room-temperature mobilities, $\mu = 26 \pm 3$ $\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ for C_{60} and 6 ± 1 $\text{cm}^2 \text{V}^{-1} \text{s}^{-1}$ for C_{70} , are 2–5 times higher than those reported previously for fullerenes in transistors^{11,12}. In general, mobilities in bulk semiconductors are larger than those at interfaces, owing to interfacial traps^{19,20}, and mobilities and diffusivities in disordered semiconductors typically increase with charge density because of the filling of shallow traps and low-energy states in the conduction-band tail²¹. The devices measured here are therefore nearly ideal for achieving high mobility, as charge transport occurs in the bulk of the channel with electron densities of more than 10^{17}cm^{-3} .

The diffusivities of DTDCTB-(10 nm C_{60})-neat and DTDCTB-(10 nm C_{70})-neat devices are thermally activated, with activation energies

$E_A = 70 \pm 8$ meV and $E_A = 36 \pm 3$ meV, respectively (see Fig. 3e, f). The exponential decrease in D with temperature is confirmation that conduction is limited by intermolecular hopping²², even though mobilities as high as those measured here are often attributed to band-like transport. In the C_{70} device, k decreases as a function of temperature ($E_A = 50 \pm 11$ meV) with nearly the same activation energy as D , suggesting that its L_D is approximately the mean free path between collisions with sparsely distributed defects. In the C_{60} devices, best fits give $k < 0.1 \text{s}^{-1}$, which suggests that the electron lifetime is longer than the timescale of the transient measurements. The values of D and k for C_{60} devices, therefore, cannot be used to accurately predict L_D . This is consistent with the fact that the total charge collected in the transient measurements does not decrease systematically with distance. However, we estimate an error of 25% in the amount of total charge collected over a distance of 1 cm, and thus a lower bound on diffusion length can be calculated using $\exp(-1 \text{ cm}/L_D) = 0.75$ (where L_D is in cm), which yields $L_D > 3.5$ cm.

We have thus demonstrated centimetre-scale electron diffusion in a photoactive, fullerene-based heterostructure, with room-temperature diffusivities of $D = 0.67 \pm 0.06 \text{cm}^2 \text{s}^{-1}$ for C_{60} and $D = 0.16 \pm 0.01 \text{cm}^2 \text{s}^{-1}$ for C_{70} , and with thermal activation energies of $E_A = 70 \pm 8$ meV and $E_A = 36 \pm 3$ meV, respectively. Among the structures explored, long-range diffusion was observed only when using d-a-a' dipolar donors in the photoactive D/A heterojunction adjacent to the electron-conducting channel. The highly dipolar donors destabilize the HOMO energy of C_{70} by 0.42 ± 0.1 eV in the blends, thereby providing energetic confinement of electrons in the channel. Monte Carlo simulations aided in understanding these results by confirming that even the lower bound of this measured energy barrier (0.32 eV) is sufficient to support centimetre-scale diffusion in the channel. The surprisingly long diffusion lengths suggest the nearly total absence of recombination centres at interfaces or within the conducting fullerene channels, even though the materials form disordered films. These results may prove useful when applied in devices in which long-range charge transport is required. For example, channel currents may open up the possibility of organic optoelectronic devices with unique properties, such as semi-transparent photovoltaics with large-period metal grid cathodes, organic field-effect transistors²³ and lateral photovoltaics²⁴. Additionally, energetically confined channels suggest that Hall-effect and lateral time-of-flight experiments are possible for the accurate characterization of organic materials. However, the presence of such channels can also result in anomalously high short-circuit currents during photovoltaic operation. In this circumstance, care must be taken to prevent the overestimation of solar cell efficiency, particularly for small-area devices if the active area is overfilled by the illumination source.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 17 June; accepted 9 November 2017.

Published online 18 January 2018.

- Coropceanu, V. *et al.* Charge transport in organic semiconductors. *Chem. Rev.* **107**, 926–952 (2007).
- Bässler, H. Charge transport in disordered organic photoconductors. *Phys. Status Solidi B* **175**, 15–56 (1993).
- Sarkar, D. & Halas, N. J. Dember effect in C_{60} thin films. *Solid State Commun.* **90**, 261–265 (1994).
- Tripathi, A. K., Tripathi, D. C. & Mohapatra, Y. N. Simultaneous and direct measurement of carrier diffusion constant and mobility in organic semiconductors and deviation from standard Einstein relation. *Phys. Rev. B* **84**, 041201 (2011).
- Xue, J., Rand, B. P., Uchida, S. & Forrest, S. R. A hybrid planar-mixed molecular heterojunction photovoltaic cell. *Adv. Mater.* **17**, 66–71 (2005).
- Xiao, X., Bergemann, K. J., Zimmerman, J. D., Lee, K. & Forrest, S. R. Small-molecule planar-mixed heterojunction photovoltaic cells with fullerene-based electron filtering buffers. *Adv. Energy Mater.* **4**, 1301557 (2014).
- Griffith, O. L. *et al.* Charge transport and exciton dissociation in organic solar cells consisting of dipolar donors mixed with C_{70} . *Phys. Rev. B* **92**, 085404 (2015).
- Ting, H. C. *et al.* Benzochalcogenodiazole-based donor-acceptor-acceptor molecular donors for organic solar cells. *ChemSusChem* **7**, 457–465 (2014).

9. Cheyns, D., Kim, M., Verreert, B. & Rand, B. P. Accurate spectral response measurements of a complementary absorbing organic tandem cell with fill factor exceeding the subcells. *Appl. Phys. Lett.* **104**, 093302 (2014).
10. Nunomura, S., Che, X. & Forrest, S. R. Charge trapping in mixed organic donor–acceptor semiconductor thin films. *Adv. Mater.* **26**, 7555–7560 (2014).
11. Itaka, K. *et al.* High-mobility C₆₀ field-effect transistors fabricated on molecular-wetting controlled substrates. *Adv. Mater.* **18**, 1713–1716 (2006).
12. Anthopoulos, T. D. *et al.* High performance *n*-channel organic field-effect transistors and ring oscillators based on C₆₀ fullerene films. *Appl. Phys. Lett.* **89**, 213504 (2006).
13. Kwiatkowski, J. J., Frost, J. M. & Nelson, J. The effect of morphology on electron field-effect mobility in disordered C₆₀ thin films. *Nano Lett.* **9**, 1085–1090 (2009).
14. Liu, X., Ding, K., Panda, A. & Forrest, S. R. Charge transfer states in dilute donor–acceptor blend organic heterojunctions. *ACS Nano* **10**, 7619–7626 (2016).
15. Bürgi, L., Friend, R. H. & Sirringhaus, H. Formation of the accumulation layer in polymer field-effect transistors. *Appl. Phys. Lett.* **82**, 1482–1484 (2003).
16. Jarrett, C. P., Pichler, K., Newbould, R. & Friend, R. H. Transport studies in C₆₀ and C₆₀/C₇₀ thin films using metal–insulator–semiconductor field-effect transistors. *Synth. Met.* **77**, 35–38 (1996).
17. Haddon, R. C. C₇₀ thin film transistors. *J. Am. Ceram. Soc.* **118**, 3041–3042 (1996).
18. Roichman, Y. & Tessler, N. Generalized Einstein relation for disordered semiconductors—implications for device performance. *Appl. Phys. Lett.* **80**, 1948–1950 (2002).
19. Sakanoue, T. & Sirringhaus, H. Band-like temperature dependence of mobility in a solution-processed organic semiconductor. *Nat. Mater.* **9**, 736–740 (2010).
20. Jurchescu, O. D., Popinciuc, M., Van Wees, B. J. & Palstra, T. T. M. Interface-controlled, high-mobility organic transistors. *Adv. Mater.* **19**, 688–692 (2007).
21. Leijtens, T., Lim, J., Teuscher, J., Park, T. & Snaith, H. J. Charge density dependent mobility of organic hole-transporters and mesoporous TiO₂ determined by transient mobility spectroscopy: implications to dye-sensitized and organic solar cells. *Adv. Mater.* **25**, 3227–3233 (2013).
22. Tummala, N. R., Zheng, Z., Aziz, S. G., Coropceanu, V. & Brédas, J.-L. Static and dynamic energetic disorders in the C₆₀, PC₆₁BM, C₇₀, and PC₇₁BM fullerenes. *J. Phys. Chem. Lett.* **6**, 3657–3662 (2015).
23. Torricelli, F., Colalongo, L., Raiteri, D., Kovács-Vajna, Z. M. & Cantatore, E. Ultra-high gain diffusion-driven organic transistor. *Nat. Commun.* **7**, 10550 (2016).
24. Kim, M. *et al.* Lateral organic solar cells with self-assembled semiconductor nanowires. *Adv. Energy Mater.* **5**, 1401317 (2015).

Acknowledgements This work was supported by the United States Department of Energy SunShot Program under awards DE-EE0006708 and DE-EE0005310, and the Air Force Office of Scientific Research under award FA9550-14-1-0245. We thank M. Ware for discussions regarding numerical simulations.

Author Contributions Q.B. fabricated all samples, performed all measurements, and assisted with data fitting and analysis. C.C. assisted with current measurements, formulated theory, and performed simulations and data analysis. X.C. analysed mismatch in organic photovoltaic devices between over- and under-filled device illumination measurements. A.P. assisted with ultraviolet photoelectron spectral measurement of film energy levels. Y.Q. calculated lateral electric field strength in the channel. S.R.F. supervised the project and analysed data. Q.B., C.C. and S.R.F. wrote the manuscript together.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.R.F. (stevefor@umich.edu).

Reviewer Information *Nature* thanks N. Banerji and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Materials and device fabrication. The materials C₇₀, C₆₀, DTDCPB, DTDCTB, SubPc, DBP and BPhen were purchased commercially and purified via sublimation before device fabrication. Glass substrates were cut into 5 cm × 5 cm squares containing four pre-patterned indium tin oxide (ITO) strips (150 nm × 1 cm × 2.3 cm) as bottom electrodes. Substrates were cleaned by sequential sonications in Tergitol, deionized water, acetone and isopropanol followed by snow-cleaning²⁵ on a 100 °C hotplate with a stream of gaseous CO₂ and a 10-minute ultraviolet–ozone exposure. The photoactive device structure was 150 nm ITO/10 nm MoO₃ anode buffer/80 nm donor:C₇₀ 1:1/fullerene/electron-blocking layer/100 nm Ag. The donor in the heterojunction was DTDCPB, DTDCTB, SubPc or DBP. The fullerene is either C₆₀ or C₇₀, and the electron blocker is either an 8-nm-thick layer of BPhen or a 10-nm-thick layer of 1:1 BPhen:C₆₀ with a 5-nm-thick BPhen cap. Mixed blocking layers are commonly used in OPVs to improve stability and electron extraction efficiency^{6,26}.

Organic materials, metals and MoO_x were deposited at rates of 0.5–1 Å s^{−1} through shadow masks in a vacuum thermal evaporator with a base pressure of 10^{−7} torr. Organics and MoO₃ were deposited through a large square mask, leaving 1 mm of the substrate uncoated around the periphery for electrical contacts to the ITO. Using a separate shadow mask, 1-mm-wide strips of 100-nm-thick silver films were deposited across the organics to form the cathode, with 0.1 cm² of intersection perpendicular to the ITO bottom contact, as shown in Fig. 1a. The steady-state photocurrent device was encapsulated using a glass cover sealed to the substrate with a bead of ultraviolet cured epoxy around its periphery.

The electrical injection device had the structure p-type Si (20 Ω per square, *h*)/500 nm SiO₂/10 nm BPhen/50 nm C₆₀/50 nm Ag/20 nm DTDCPB:C₇₀ 1:1/8 nm BPhen/100 nm Ag. The 50-nm-thick silver layer (collecting contact), and the 100-nm-thick silver layer (injecting contact) were patterned into 1 mm × 18 mm parallel strips using a shadow mask, separated by 1 cm. The substrate oxide was etched, using buffered HF, to allow contact with the silicon in a small area away from the device. For alignment purposes, two electrically inactive silver strips positioned above the collecting contact were deposited simultaneously with the injecting contact. We used the same heterojunction/channel/electron-blocking layer structure to enable a direct comparison between the electrical and optical devices, although long-range diffusion is expected with any trap-free blocking interface.

Transient photocurrent measurements. Transient photocurrent measurements were performed under vacuum in an open-loop liquid N₂ cryostat with four vacuum feedthroughs for electrical contacts, optical fibre and micrometer positioning arm. The fibre, which had a 25-μm-diameter core with a numerical aperture of 0.1, and a measured Gaussian output width of <40 μm, was positioned normal to the device surface. Pulses (duration of 0.5 ms to 2 ms and delays of 1 s to 100 s between pulses) from diode lasers (wavelengths 405 nm and 637 nm) were focused into the fibre for temporal measurements. The position of the fibre was controlled with an *x*–*y*–*z* micrometer positioning stage. A 99% optically absorptive black foil was placed beneath the device to minimize light scattering. The current response was amplified with a low-noise current amplifier at 10⁸ V A^{−1} and recorded with a digital oscilloscope. Rise time filters between 10 μs and 10 ms were used on the current amplifier to minimize noise, and all spectra were averaged over at least 10 pulses. Temperature was controlled by the liquid N₂ flow rate and a resistive heater, and monitored with a thermocouple. The steady-state dark current was subtracted, leaving only the transient response to the light pulses.

Transient current measurements. Transient current measurements of the electrical injection device were performed in the dark, under vacuum and at room temperature. The silicon substrate was grounded, the injecting contact was connected to a pulse generator, and the current transient at the collecting contact was measured with a current amplifier (10⁷ V A^{−1}, 10-ms-risetime filter) and digital oscilloscope.

Ultraviolet photoelectron spectroscopy. The HOMO energies were measured using ultraviolet photoelectron spectra taken in high vacuum (10^{−8} torr) with a 21.2 eV photon source. Organic thin films were grown on conductive ITO substrates as described above. The lowest unoccupied molecular orbital energies were estimated using the low energy optical absorption edge of the material. The energy barrier between C₇₀ in the donor:C₇₀ blend and the neat C₆₀ layer on top was estimated by measuring the binding energy of the C₇₀ HOMO in a 10-nm-thick donor:C₇₀ blend with a 5-nm-thick C₆₀ cap, measuring the binding energy of the C₆₀ HOMO and taking the difference between these energies.

Charge-diffusion simulations. Charge dynamics were simulated from a solution to the diffusion equation

$$\dot{N}(x, y, t) = D \nabla^2 N(x, y, t) - kN(x, y, t) + G(x, y, t) \quad (1)$$

subject to blocking boundary conditions along the edges of the organic film and a quenching boundary condition at the edge of the silver cathode. Here, *N* is the electron density, *x* is the distance from the cathode, *y* is the lateral position, *t* is the time, *D* is the diffusivity, *k* is the sum of the trapping and recombination rates, and *G* is the generation rate. Initially, *N*(*x*, *y*, 0) = 0. The generation term is given by

$$G(x, y, t) = \frac{Q}{qt_{\text{pulse}}} \frac{1}{2\pi\sigma^2} \left(\frac{(x - x_0)^2 + (y - y_0)^2}{2\sigma^2} \right) \quad (2)$$

where *Q* is the charge injected into the fullerene transport layer, *σ* = 40 μm is the laser beam diameter, (*x*₀, *y*₀) and *t*_{pulse} are the position and length of the excitation, respectively. The intrinsic carrier density in the fullerene layer is typically <10⁸ cm^{−3}, depending on the purity²⁷, which is many orders of magnitude below the optically excited and electrically injected charge densities.

The current transient is given by the diffusion current into the cathode as a function of *t*:

$$I(t) = q \int D \frac{dN(x, y, t)}{dx} \bigg|_{x=0} dy \quad (3)$$

Equations (1)–(3) are solved numerically, and equation (3) is fitted to the data, with parameters *D*, *k* and *Q*. The diffusivity primarily determines the arrival time of the current pulse and the slope of the falling edge, *k* determines the slope at long times, and *Q* scales linearly with amplitude.

Because of the device symmetry and the blocking conditions on the perimeter of the organic films, diffusion parallel to the cathode interface does not affect the arrival time of electrons at the cathode except in the case of diffusion around a cut in the organic film. The simulated geometry is therefore one-dimensional in most cases, simplifying computation.

Diffusion simulations successfully fitted the data in Figs 1, 3 and Extended Data Fig. 1, except at the leading edge of the current transient. During measurement, some scattered light is absorbed along the channel between the silver edge and the intended point of photoexcitation, generating a prompt current response that is not accounted for by the model. This response is delayed with increasing *L*, as can be seen in Fig. 1b, as the scattered light is absorbed farther from the cathode. Bandwidth filters were used to reduce noise on the low-amplitude spectra at larger *L*. Scattered light effects were considerably suppressed by placing a highly absorptive foil beneath the device during all measurements. The resistance-capacitance time constant of the measurement circuit was <100 ns and the amplifier rise-time was <2 μs; both are many orders of magnitude shorter than the current response from the samples.

The falling edge of the electrical injection device transient was also simulated using this method, in which electrons were generated at a constant rate beneath the injecting contact until a steady-state current was reached at the collecting electrode. The collecting contact was placed at *x* = 0, with the edge of the injecting contact at *x* = 1 cm.

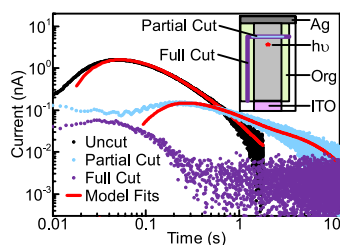
Monte Carlo simulations. Monte Carlo simulations of charge diffusion were performed on a simple cubic lattice that contained 200 × 7 × 100 sites in the *x*, *y* and *z* dimensions, respectively, with lattice constant *a* = 1 nm. Here, *x* and *y* are the directions parallel to the substrate, and *z* extends vertically from the substrate. Sites with *z* < 90*a* were designated as the donor:C₇₀ blend, and the top 10 sites represented the neat fullerene channel. The blend was randomly generated, with 50% of the sites occupied by fullerenes, corresponding to a 1:1 blend. Periodic boundary conditions in *y* were used, with blocking interfaces at *x* = 0 (the periphery of the organic films) and *z* = 100*a* (the interface between the fullerene channel and the electron-blocking layer). Quenching boundary conditions were assumed at *z* = 0 (the MoO₃/heterojunction interface) and *x* = 200*a* to collect electrons. Charges originating at (1, 1, 1) were allowed to diffuse using the Miller–Abrahams hopping rate, as follows: during each step of the computation, a random direction was chosen for an electron hopping attempt, with the probability of success given by *p* = exp(−*E*/*k*_B*T*), where *E* is the energy difference between the two sites. We let *E* = *E*_B for hops from the channel to the blend, *E* = ∞ for hops onto donors, and *E* = 0 otherwise. This assumes that the barrier for intermolecular hopping can be neglected in the lateral diffusion efficiency calculations as it does not affect the relative probability of hopping over the barrier versus laterally. Energetic disorder, which may decrease the calculated value of *L*_D, is also assumed to be small compared with the *E*_B. The charge-diffusion efficiency over the length of the simulated lattice, *η*_D[′], is given by the ratio of charges quenched at *x* = 200*a* versus at *z* = 0. Thus, the diffusion efficiency over a distance, *x*, is given by: *η*_D(*x*) = (*η*_D[′])^{*x*/*d*}, where *d* is the length of the simulated lattice and the charge-diffusion length is: *L*_D = −*d*/ln(*η*_D[′]).

Photovoltaic characterization. Current–voltage characteristics were measured using a semiconductor parameter analyser both in the dark and under 100 mW cm^{−2} of AM1.5G filtered illumination from a xenon arc lamp. External

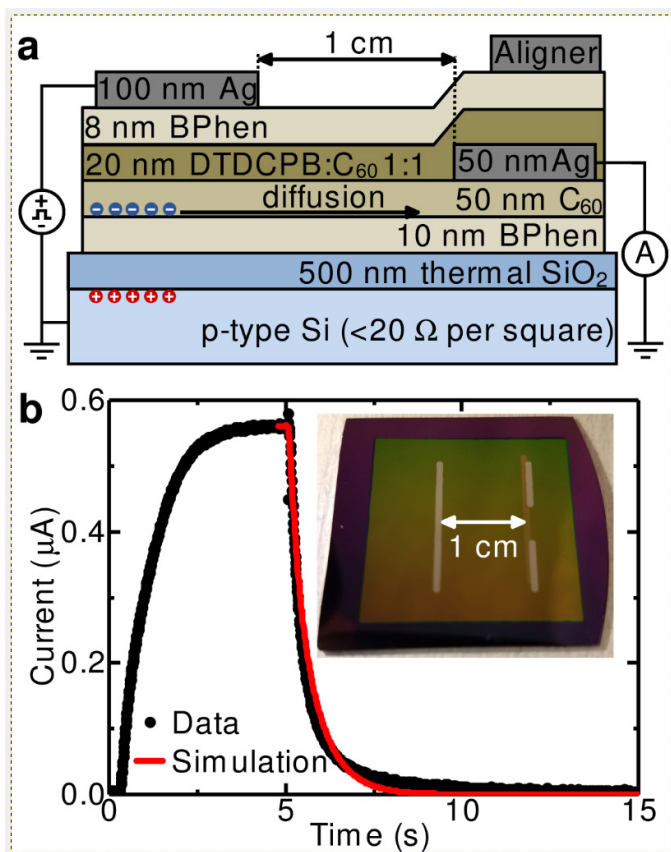
quantum efficiency was measured with a lock-in current amplifier in response to a calibrated monochromated light source chopped at 100 Hz. The steady-state device response to 633-nm illumination was measured using an identical set-up to that described for transient photocurrent measurements with a 10 mW He–Ne laser focused into the fibre.

Data availability. All relevant data are available from the corresponding author on reasonable request.

25. Wang, N. *et al.* Snow cleaning of substrates increases yield of large-area organic photovoltaics. *Appl. Phys. Lett.* **101**, 133901 (2012).
26. Burlingame, Q. *et al.* Reliability of small molecule organic photovoltaics with electron-filtering compound buffer layers. *Adv. Energy Mater.* **6**, 1601094 (2016).
27. Olthof, S. *et al.* Ultralow doping in organic semiconductors: evidence of trap filling. *Phys. Rev. Lett.* **109**, 176601 (2012).



Extended Data Figure 1 | Impact of channel disruption on channel currents. Room-temperature transient currents were measured on a DTDCTB-(10 nm C_{60})-mixed device at $L = 3$ mm before and after a series of razor blade cuts was made to the organic layers, as shown in the inset. The peak height of the current pulse was greatly reduced and the peak arrival time was delayed for devices with a ‘partial cut’ that was transverse to, and spanned the width of, the ITO anode between the illumination position and the silver cathode, compared with the pristine, uncut device. Charge-diffusion simulations were performed for both geometries, where the only difference was a blocking boundary condition at the position of the partial cut. We find that charge diffusion around the cut accounts for the differences between the cut and uncut device transients, as demonstrated by the agreement between fits (lines) and the data. The partial cut was also extended such that there was no continuous organic path between the illumination position and the cathode (a ‘full cut’ device). This eliminated the response except for a residual current at time $t < 200$ ms arising from scattered light absorbed in the organic layers between the cathode and cut. This effect was observed in all devices exhibiting channel currents.



Extended Data Figure 2 | Electron diffusion in an electrical injection device. **a**, A device was fabricated to characterize charge diffusion in an electron-only electrically injected channel, with the structure shown here. Charges were injected into the C_{60} channel by applying a 50 V pulse between the injecting contact and the silicon substrate for 5 s. **b**, The transient current collected at the buried contact. A steady-state current of $0.56 \mu\text{A}$ is observed approximately 3 s after the start of the pulse, with an exponential decay time of about 400 ms. A simulation of the turn-off transient using the same optically measured parameters for D and k in Table 1 for the DTDCPB-(10 nm C_{60})-neat device is also shown in **b** (solid line). The small deviations of the fits to the electrical data are probably due to slow de-trapping of charges in the BPhen and SiO_2 that are injected during the 50 V pulse. Inset, photograph of the device with 1 cm scale bar.

Small-scale soft-bodied robot with multimodal locomotion

Wenqi Hu^{1*}, Guo Zhan Lum^{1*}, Massimo Mastrangeli¹ & Metin Sitti¹

Untethered small-scale (from several millimetres down to a few micrometres in all dimensions) robots that can non-invasively access confined, enclosed spaces may enable applications in microfactories such as the construction of tissue scaffolds by robotic assembly¹, in bioengineering such as single-cell manipulation and biosensing², and in healthcare^{3–6} such as targeted drug delivery⁴ and minimally invasive surgery^{3,5}. Existing small-scale robots, however, have very limited mobility because they are unable to negotiate obstacles and changes in texture or material in unstructured environments^{7–13}. Of these small-scale robots, soft robots have greater potential to realize high mobility via multimodal locomotion, because such machines have higher degrees of freedom than their rigid counterparts^{14–16}. Here we demonstrate magneto-elastic soft millimetre-scale robots that can swim inside and on the surface of liquids, climb liquid menisci, roll and walk on solid surfaces, jump over obstacles, and crawl within narrow tunnels. These robots can transit reversibly between different liquid and solid terrains, as well as switch between locomotive modes. They can additionally execute pick-and-place and cargo-release tasks. We also present theoretical models to explain how the robots move. Like the large-scale robots that can be used to study locomotion¹⁷, these soft small-scale robots could be used to study soft-bodied locomotion produced by small organisms.

Our robot is constructed of soft active materials, which can be magnetically actuated to generate desired time-varying shapes¹⁶ (see Supplementary Information section S1). Although our robotic system includes both an untethered soft device and the electromagnets that remotely generate the actuating fields (see Supplementary Information section S2 and Supplementary Fig. 2), we refer to only the untethered soft device as a ‘robot’, for consistency with the literature^{3–5,16,18,19}. Unlike previous robots constructed with similar materials^{7,16}, our proposed robot design and actuation inputs can achieve multimodal locomotion, and we have concurrently accounted for the robot’s programmed soft-bodied deformation and rigid-body rotation characteristics in different terrains. The choice of magnetic actuation suits various applications because the actuating fields can easily and harmlessly penetrate most biological and synthetic materials^{3,4}. This work uses external (off-board) magnetic actuation only, but it should also be possible to create similar soft machines that use internal (on-board) soft actuation methods²⁰ to produce similar time-varying shapes and rotation.

The magneto-elastic, rectangular-sheet-shaped, soft robot is made of silicone elastomer (Ecoflex 00-10) embedded with hard magnetic neodymium-iron-boron (NdFeB) microparticles that have an average diameter of 5 μm . The surfaces of the robot are hydrophobic, and they can potentially be made biocompatible²¹ (Supplementary Information section S1C). By following the magnetization process described in Supplementary Information section S1A, the robot can be programmed to have a single-wavelength harmonic magnetization profile \mathbf{m} along its body (Fig. 1a and Supplementary Fig. 1). After \mathbf{m} is programmed,

the robot can be controlled by a time-varying magnetic field \mathbf{B} to generate different modes of locomotion. Unless otherwise specified, \mathbf{B} is spatially uniform, and therefore no magnetic forces are applied to translate the robot (Supplementary Information section S15). The uniform \mathbf{B} , however, can control the robot’s morphology and steer it to move in a desired direction. To describe the effects of \mathbf{B} , we express $\mathbf{B} = [\mathbf{B}_{xy}^T, B_z]^T$ with respect to the robot’s body frame (Fig. 1a) where \mathbf{B}_{xy} represents the x – y plane components of \mathbf{B} , that is, $\mathbf{B}_{xy} = [B_x, B_y]^T$. The interaction between \mathbf{B}_{xy} and \mathbf{m} produces spatially varying magnetic torques that deform the robot, and hence controlling \mathbf{B}_{xy} allows us to generate the desired time-varying shapes for the robot. As the deformed robot possesses an effective magnetic moment \mathbf{M}_{net} (Fig. 1b), which tends to align with \mathbf{B} , we can control B_z to rotate the robot about its y axis, steering it along a desired direction (see Supplementary Information section S3B(II)).

Depending on the magnitude of \mathbf{B}_{xy} —that is, B_{xy} —the robot exhibits different shape-changing mechanisms (Fig. 1b and Supplementary Information section S3A–B). When B_{xy} is small (for example, $< 5 \text{ mT}$) and \mathbf{B}_{xy} is aligned along the two principal directions shown in Fig. 1b (II and III), the prescribed \mathbf{m} produces a sine or a cosine shape for the robot. Because the robot’s deformation is small in such conditions, orienting \mathbf{B}_{xy} away from the principal directions generates a weighted superposition of the two basic configurations. Thus, we can create a travelling wave along the robot’s body by using a rotating \mathbf{B}_{xy} that has a small constant magnitude. As the robot’s \mathbf{M}_{net} is always parallel to the applied \mathbf{B}_{xy} in small-deflection conditions, the robot does not experience any rigid-body magnetic torque and consequent rotation about its z axis (Supplementary Information section S3B(I)). Conversely, when \mathbf{B}_{xy} has high magnitude (for example, $B_{xy} = 20 \text{ mT}$) and is aligned along the principal axis shown in Fig. 1b (IV and V), the robot undergoes a large-deflection shape change, deforming into either a ‘C’- or a ‘V’-shape. However, if the direction of \mathbf{B}_{xy} is not along this principal axis, the deformed robot generates a large \mathbf{M}_{net} that is generally non-parallel to the applied field, and this makes the robot rotate about its z axis until its \mathbf{M}_{net} aligns with \mathbf{B}_{xy} (Fig. 1c and Supplementary Information section S3B(I)). At the end of this rotation, the robot will assume its ‘C’- or ‘V’-shape configuration because the generated \mathbf{M}_{net} in these configurations is naturally aligned with the applied \mathbf{B}_{xy} . Using this mechanism, we can control the robot’s angular displacement about its z axis to enable locomotion modalities like rolling, walking and jumping.

By using the steering and shape-changing mechanisms, we demonstrate all of our robot’s locomotion modes in Figs 2 and 3. When completely immersed in water, the robot can swim upwards and overcome gravity (Fig. 2a, Supplementary Video 1, and Supplementary Information section S10). A periodic \mathbf{B} with time-varying magnitude along the principal axis allows the shape of the robot to alternate between the ‘C’- and ‘V’-shapes, enacting a gait similar to jellyfish swimming²². Inertial effects at Reynolds number ranging from 74 to 190 permit this time-symmetric but speed-asymmetric swimming gait

¹Physical Intelligence Department, Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany.

*These authors contributed equally to this work.

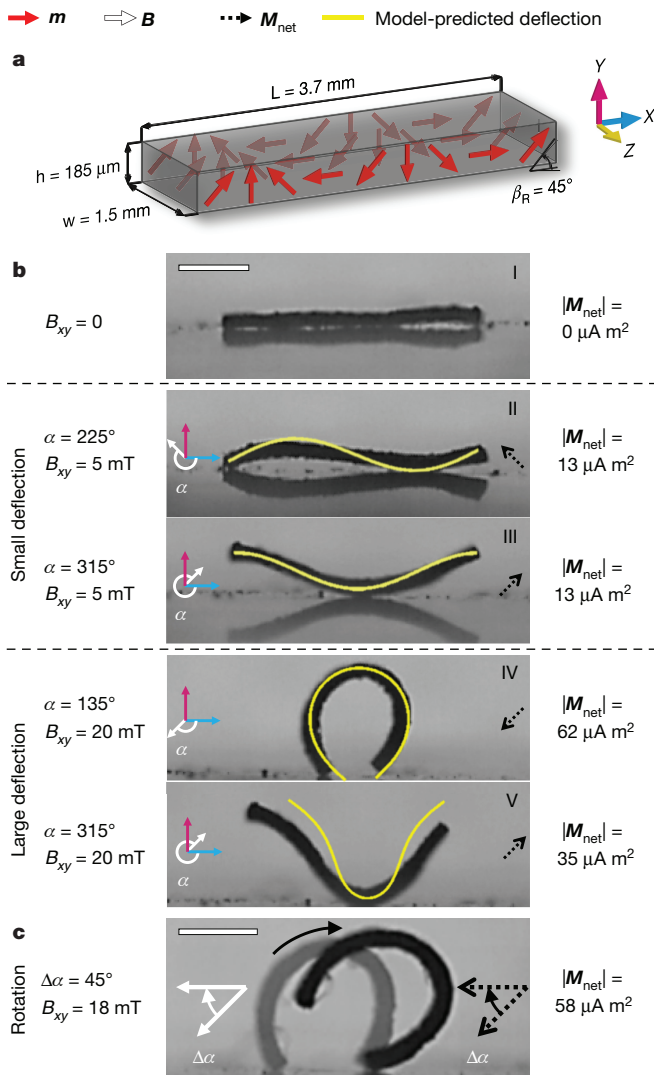


Figure 1 | Design, shape-change mechanisms and rigid-body rotation of the magneto-elastic soft millirobot. **a**, The rectangular-sheet-shaped magnetic soft robot's dimensions and magnetization profile m . The profile can be described as a single-wavelength harmonic function for its m_x and m_y components along the robot's length L and equally along its width w and thickness h . β_R is the phase shift in m that determines the principal directions (Supplementary Information section S3A). **b**, Theoretically predicted (Supplementary Information section S3A) and experimental shapes of a magnetically deformed millirobot. α is a clockwise angle from the x axis, and it is used for describing the direction of B_{xy} . In (I), the robot is in its rest state under null B_{xy} . A small residual curvature is induced by the pre-stress caused by the demoulding process. (II) and (III) show small-deformation shapes under a small-magnitude B_{xy} ($B_{xy} = 5 \text{ mT}$) aligned along the two principal directions ($\alpha = 225^\circ$ for the sine shape and 315° for the cosine shape) in the x - y plane. In contrast, the robot deforms into a 'C'-shape (IV) or 'V'-shape (V) when a large-magnitude B_{xy} ($B_{xy} = 20 \text{ mT}$) aligns along the principal axis ($\alpha = 135^\circ$ and 315° , respectively). The induced pre-stress from the demoulding process may introduce a deviation between the predicted and actual shapes, particularly for (II) and (V). The theoretical M_{net} of the robot for each shape is shown on the right. **c**, The clockwise rotation of the 'C'-shaped robot is induced by rotating B_{xy} . Scale bars, 1 mm.

to produce fluid vortices that propel the robot to the water surface (Fig. 3c, Supplementary Video 1, and Supplementary Fig. 37). Upon emersion, the soft robot strongly pins at the water–air interface by exposing its hydrophobic surface to air.

Inspired by beetle larva that overcome frictionless barriers by performing quasi-static work on liquid–air interfaces²³, the soft robot

can climb up a water meniscus by deforming into a 'C'-shape to enhance its liquid buoyancy without extra energy expenditure (Fig. 2b, Supplementary Video 2 and Supplementary Information section S7). Upon meniscus climbing and reaching contact with an adjacent solid platform, a slow rotating B will make the 'C'-shaped robot rotate about its z axis. The hydrophobicity of its surface allows the robot to be peeled away from the water surface by such rotation (Fig. 2c, Supplementary Video 2 and Supplementary Information section S11B). In contrast to meniscus climbing, the robot can also dive into the liquid bulk by disengaging from the water–air interface via a fast sequence of downward bending, rotation and flipping (Fig. 2d and Supplementary Information section S11A).

In nature, soft-bodied caterpillars use rolling locomotion to escape from their predators, because this is an efficient and fast way to sweep across solid terrains²⁴. Like caterpillars, our robots can also roll directionally over a rigid substrate or dive from a solid onto a liquid surface (Figs 2e and 3a). This locomotion is enabled by a high-magnitude rotating B (such as $B = 18.5 \text{ mT}$), which allows the robot to roll in its 'C'-shape configuration (Supplementary Video 3 and Supplementary Information section S5). However, the curled-up robot cannot roll across substrate gaps wider than its diameter but narrower than the length of the robot; such gaps can instead be traversed by walking.

Walking is a particularly robust way to move over unstructured surfaces and affords precise tuning of stride length and frequency (Fig. 2f, Supplementary Video 3 and Supplementary Information section S6). Inspired by the walking gait of inchworms²⁵, the robot can walk in a desired direction when we use a periodic B to sequentially adapt its tilting angle and curvature. In each walking cycle, the robot first anchors on its front end to tilt forward so that it can pull its back end forward. The robot then anchors on its back end to tilt backwards and extends its front end to achieve a positive stride in a single cycle.

When the walking robot is blocked by narrow openings, it can mimic another caterpillar locomotion²⁴ and use an undulating gait to crawl through the obstacle (Fig. 2g, Supplementary Video 4 and Supplementary Information section S9). Crawling is encoded by a rotating B to produce a longitudinal travelling wave that propels the robot along the direction of the wave. A similar control sequence additionally enables the robot to produce an undulating gait to swim efficiently on liquid surfaces²⁶ (Fig. 3a and Supplementary Video 6). In contrast to crawling, however, the undulating swimming direction is antiparallel with the direction of the travelling waves. Although previous robots with multi-wavelength, harmonic magnetization profiles have also demonstrated undulating swimming locomotion⁷, such robots have not been able to create the critical 'C'- and 'V'-shapes necessary to realize multimodal locomotion.

Like nematodes²⁷, the soft robot can jump over obstacles that are too high or too time-consuming to roll or walk over, by imparting an impulsive impact on a rigid surface (Fig. 2h, Supplementary Video 5 and Supplementary Information section S4). The B control sequence prompts both the robot's rigid-body rotation, which specifies the jumping direction, and elastic deformations to maximize the momentum of its free ends before striking the substrate. This sequence of B is specified in the robot's local y - z plane, where B_y is used for inducing the shape-changing mechanism, whereas the rigid-body rotation of the robot is induced by both B_y and B_z .

To illustrate the robot's potential to navigate across unstructured environments (Supplementary Information section S13), we demonstrate that the robot can use a series of locomotion modes to fully explore a hybrid liquid–solid environment (Fig. 3 and Supplementary Video 6) and a surgical human stomach phantom (Fig. 4a, Supplementary Video 7 and Supplementary Information section S14A). Heading towards an *in vivo* ultrasound-guided operation, we also show that the robot can be visualized by an ultrasound medical imaging device as it rolls within the concealed areas of *ex vivo* chicken muscle tissue (Fig. 4b, Supplementary Video 8, Supplementary Information section S14B and Supplementary Fig. 44). The soft robot can additionally

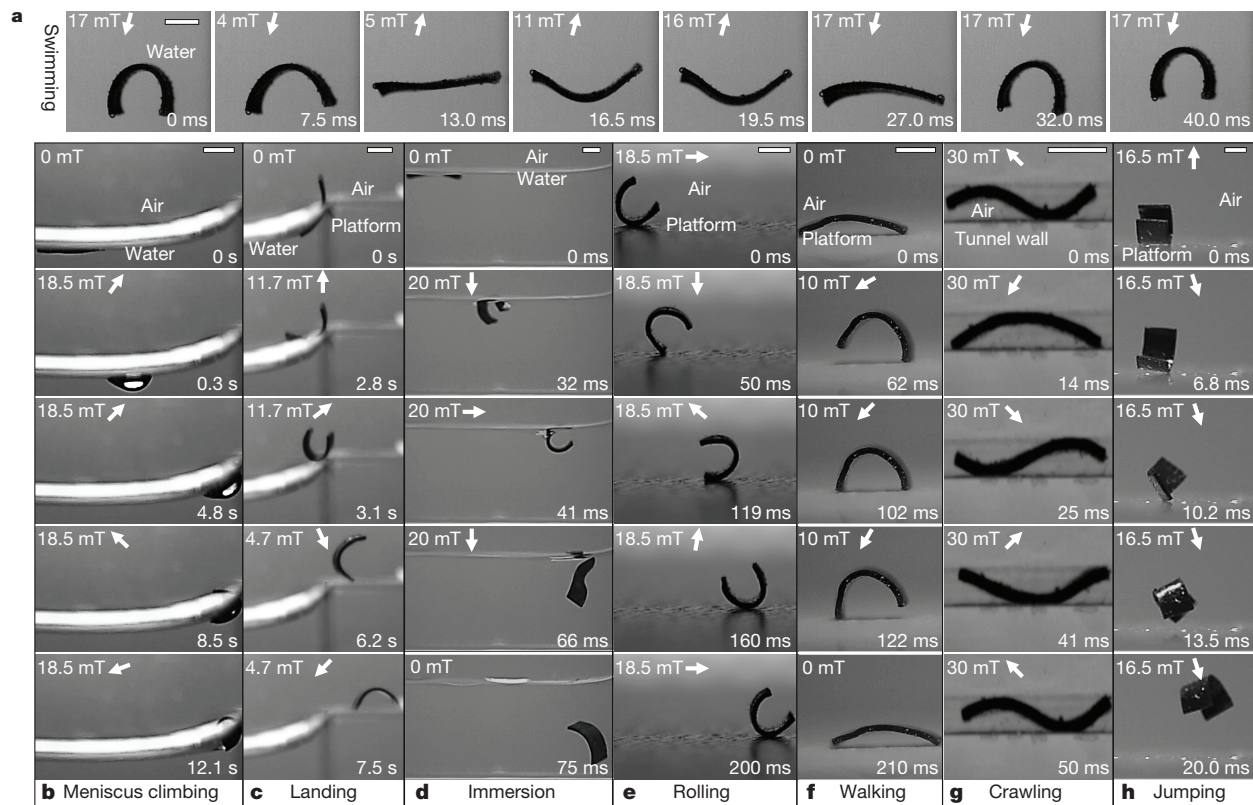


Figure 2 | Locomotion and transition modes of the soft millirobot. The sub-panels for the swimming locomotion in **a** are displayed in a horizontal sequence to show the relative vertical displacement between them. The sub-panels for the other types of locomotion are arranged in a vertical sequence to show the relative horizontal displacement between them. The input (**B**) sequences for each locomotion mode are detailed in Supplementary Information sections S4–S11. The corresponding time stamps and **B** in each sub-panel are shown in the bottom-right and top-left corners, respectively. **a**, Jellyfish-like swimming in water using a time-symmetric but speed-asymmetric gait. **b**, Water meniscus climbing. An anticlockwise magnetic torque progressively adapts the pose of the robot as it deforms and ascends owing to buoyancy. **c**, Landing, that is,

transition from water surface onto solid ground. A clockwise-rotating **B** peels the robot off the water surface and lets it stand on the platform. **d**, Immersion, that is, transition from the surface into the bulk of a water pool by a combination of curling and rigid-body rotation. **e**, Rolling by a clockwise-rotating **B** of high magnitude. The robot tilts and changes its curvature to create a net stride in each cycle. **f**, Walking. The robot tilts and changes its curvature to create a net stride in each cycle. **g**, Crawling inside a tubular channel with a cross-section of $0.645 \text{ mm} \times 2.55 \text{ mm}$ by using an undulating travelling wave along the robot's body. **h**, Directional jumping. The robot uses its rigid-body motion and shape change to induce a jumping momentum. More than one robot is used for this illustration but all of these robots have the same design. Scale bars, 1 mm.

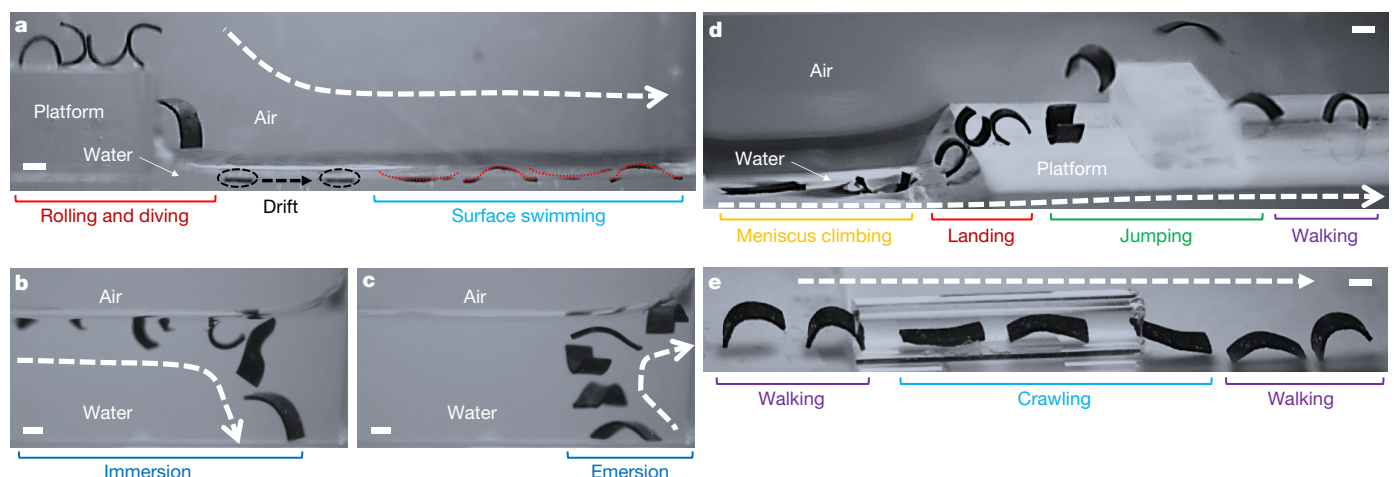


Figure 3 | Multimodal locomotion over a hybrid liquid–solid environment. **a**, The soft robot rolls and dives from a solid platform into the adjacent water pool, where it drifts away along the water meniscus. The undulating robot then swims rightwards. **b**, **c**, The robot rotates, disengages from the water surface, sinks, and subsequently swims up from the pool bottom to emerge again at the water–air interface. **d**, The robot climbs up a water meniscus, lands on the solid platform, jumps beyond a

standing obstacle, and walks away. **e**, The robot walks towards a tubular tunnel (diameter 1.62 mm) that impedes its walking gait. The robot then switches to the crawling mode to cross the tunnel, and finally walks away. The locomotion modes were sequentially captured in four separate videos owing to the restrictions of the workspace (Supplementary Information section S2A). Only one robot is used in this illustration. Scale bars, 1 mm.

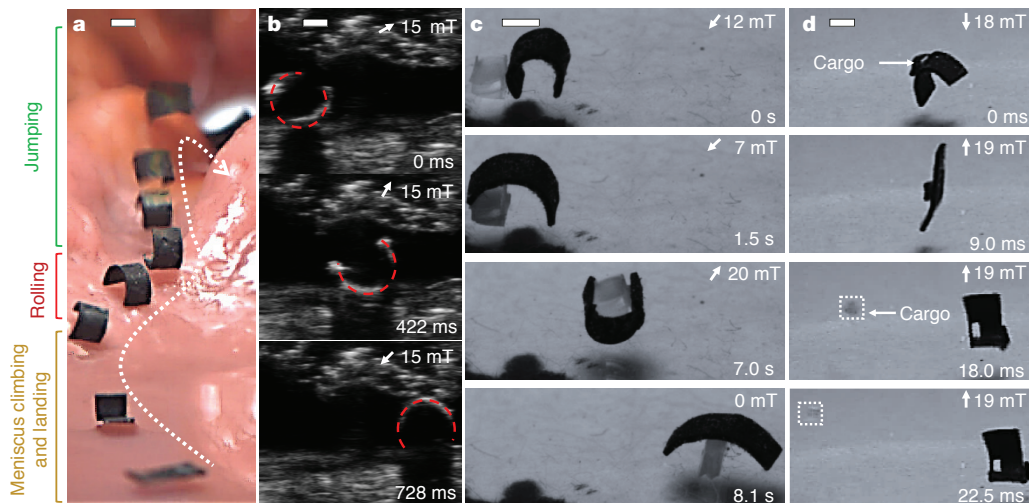


Figure 4 | Towards medical applications. **a**, The soft robot navigating across a synthetic stomach phantom using a combination of locomotion modes (Supplementary Video 7). **b**, Ultrasound-guided locomotion: the robot (marked out by dotted red lines) rolls within the concealed areas of an *ex vivo* chicken tissue, as visualized by ultrasound imaging (Supplementary Video 8 and Supplementary Information section S14B). **c**, The robot approaches a cargo item (nylon, 1 mm × 0.8 mm × 1.5 mm) by walking on a flat rigid surface, picks up the cargo by curling into the

'C'-shape, transports away the cargo by rolling and maintaining its 'C'-shape configuration, and releases the cargo by uncurling at a new position (Supplementary Video 9). **d**, Dynamic and selective cargo release (Supplementary Video 10). A paper tissue (0.5 mm × 0.5 mm × 0.1 mm, used as model drug container) is bound to the robot by an extra appendage (Supplementary Information section S14C). After pre-bending the robot, **B** is quickly reversed to open the appendage and release the cargo (cargo highlighted by the white dashed square). Scale bars, 1 mm.

accomplish functional tasks like gripping an object and transporting it to a targeted location (Fig. 4c and Supplementary Video 9), as well as ejecting a cargo that is strapped onto the robot (Fig. 4d, Supplementary Video 10 and Supplementary Information section S14C).

In addition to magnetic field-induced torques, magnetic gradient-based pulling forces could also be used to enhance locomotion performance (for example, speed and jumping height). Moving along this direction, we show that the jumping height can be increased by adding magnetic gradient-based pulling forces (Supplementary Video 5), and we will explore other similar possibilities in the future. Using gradient-based pulling exclusively may however be detrimental, as the dynamics of this actuation method is inherently unstable²⁸. From a practical standpoint, gradient-based pulling methods are also less energy efficient than locomotion propelled via magnetic field-based torques²⁹ (Supplementary Information section S15).

The lack of an on-board actuation method prevents the proposed robot from operating in large open spaces, making it unsuitable for outdoor applications such as environment exploration and monitoring. Furthermore, the current demoulding process creates a pre-stress in the magneto-elastic material that induces a small residual curvature in the robot when it is in the rest state (shape I in Fig. 1b and Supplementary Information section S1D). Although the pre-stress does not hinder the robot from achieving multiple modes of locomotion and could be reduced through improved fabrication, it induces small errors in the predicted robot shapes (shapes II and V in Fig. 1b) and partly affects the model matching of the experimental data for the walking and undulating swimming speeds (Supplementary Information sections S6 and S8).

To understand small-scale soft-bodied robot locomotion better, we devised theoretical models to perform a scaling analysis on how the robot's dimensions (L , w and h , shown in Fig. 1a) would affect the jumping, rolling, walking, meniscus-climbing and undulating swimming locomotion modalities (Supplementary Information sections S4–S8). The theoretical models for the crawling and jellyfish-like swimming locomotion are too difficult to be derived, so we instead used the experimental data in Supplementary Information sections S9–S10 to derive corresponding fitting models. From our theoretical and fitting models, we predict that a larger L and a smaller h are always preferred for multimodal locomotion because a longer

and thinner rectangle shape helps the robot to move faster and jump higher. The models also suggest that w would affect only the jellyfish-like swimming locomotion and that minimizing w would increase the swimming speed. There are, however, practical lower bounds for both h and w , because our current fabrication technique has difficulties in creating robots that have $h < 40 \mu\text{m}$ and $w < 0.3 \text{ mm}$. Likewise, the practical upper bound of L is typically constrained by the size requirements of specific applications and the maximum allowable workspace of the electromagnetic coil setup that generates the spatially uniform **B**. A more detailed summary for the scaling analysis and fabrication limits can be found in Supplementary Information section S12.

To validate the theoretical models, we compared them against extensive experimental characterizations conducted across robots with differing dimensions. In general, except for the undulating swimming locomotion, the experimental data agree well with our models (see Supplementary Information sections S4–S8, S12 and Supplementary Table 4). Detailed discussions pertaining to the theoretical and experimental discrepancy for the undulating swimming locomotion can be found in Supplementary Information section S8. These analyses may also provide useful design guidelines for optimizing the performance of future miniature robots that have multimodal locomotion.

We intend to use our robot to study small-scale soft-bodied locomotion on other complex terrains such as within non-Newtonian fluids and on granular media³⁰. We also plan to scale down the robots to the sub-millimetre scale and to investigate their potential *in vivo* medical applications.

Data Availability All data generated or analysed during this study are included in the published article and its Supplementary Information, and are available from the corresponding author on reasonable request.

Received 19 September; accepted 1 December 2017.

Published online 24 January 2018.

- Chung, S. E., Dong, X. G. & Sitti, M. Three-dimensional heterogeneous assembly of coded microgels using an untethered mobile microgripper. *Lab Chip* **15**, 1667–1676 (2015).
- Ceylan, H., Giltinan, J., Kozielski, K. & Sitti, M. Mobile microrobots for bioengineering applications. *Lab. Chip* **17**, 1705–1724 (2017).

3. Nelson, B. J., Kaliakatsos, I. K. & Abbott, J. J. Microrobots for minimally invasive medicine. *Annu. Rev. Biomed. Eng.* **12**, 55–85 (2010).
4. Sitti, M. *et al.* Biomedical applications of untethered mobile milli-/microrobots. *Proc. IEEE* **103**, 205–224 (2015).
5. Sitti, M. Miniature devices: voyage of the microrobots. *Nature* **458**, 1121–1122 (2009).
6. Sitti, M. *Mobile Microrobotics* (MIT Press, 2017).
7. Diller, E., Zhuang, J., Lum, G. Z., Edwards, M. R. & Sitti, M. Continuously distributed magnetization profile for millimeter-scale elastomeric undulatory swimming. *Appl. Phys. Lett.* **104**, 174101 (2014).
8. Huang, H. W., Sakar, M. S., Petruska, A. J., Pané, S. & Nelson, B. J. Soft micromachines with programmable motility and morphology. *Nat. Commun.* **7**, 12263 (2016).
9. Maeda, S., Hara, Y., Sakai, T., Yoshida, R. & Hashimoto, S. Self-walking gel. *Adv. Mater.* **19**, 3480–3484 (2007).
10. Miyashita, S., Guitron, S., Luidersdorfer, M., Sung, C. R. & Rus, D. An untethered miniature origami robot that self-folds, walks, swims, and degrades. *IEEE Int. Conf. on 'Robotics and Automation'* 1490–1496, <http://ieeexplore.ieee.org/document/7139386/> (Institute of Electrical and Electronics Engineers (IEEE), 2015).
11. Koh, J. S. *et al.* Jumping on water: surface tension-dominated jumping of water striders and robotic insects. *Science* **349**, 517–521 (2015).
12. Yuk, H., Kim, D., Lee, H., Jo, S. & Shin, J. H. Shape memory alloy-based small crawling robots inspired by *C. elegans*. *Bioinspir. Biomim.* **6**, 046002 (2011).
13. Diller, E. & Sitti, M. Three-dimensional programmable assembly by untethered magnetic robotic micro-grippers. *Adv. Funct. Mater.* **24**, 4397–4404 (2014).
14. Rus, D. & Tolley, M. T. Design, fabrication and control of soft robots. *Nature* **521**, 467–475 (2015).
15. Wehner, M. *et al.* An integrated design and fabrication strategy for entirely soft, autonomous robots. *Nature* **536**, 451–455 (2016).
16. Lum, G. Z. *et al.* Shape-programmable magnetic soft matter. *Proc. Natl Acad. Sci. USA* **113**, E6007–E6015 (2016).
17. Aguilar, J. *et al.* A review on locomotion robophysics: the study of movement at the intersection of robotics, soft matter and dynamical systems. *Rep. Prog. Phys.* **79**, 110001 (2016).
18. Diller, E., Giltinan, J., Lum, G. Z., Ye, Z. & Sitti, M. Six-degree-of-freedom magnetic actuation for wireless microrobotics. *Int. J. Robot. Res.* **35**, 114–128 (2016).
19. Kummer, M. P. *et al.* OctoMag: an electromagnetic system for 5-DOF wireless micromanipulation. *IEEE Trans. Robot.* **26**, 1006–1017 (2010).
20. Hines, L., Petersen, K., Lum, G. Z. & Sitti, M. Soft actuators for small-scale robotics. *Adv. Mater.* **29**, 1603483 (2017).
21. Amjadi, M., Yoon, Y. J. & Park, I. Ultra-stretchable and skin-mountable strain sensors using carbon nanotubes-Ecoflex nanocomposites. *Nanotechnology* **26**, 375501 (2015).
22. Gemmell, B. J. *et al.* Passive energy recapture in jellyfish contributes to propulsive advantage over other metazoans. *Proc. Natl Acad. Sci. USA* **110**, 17904–17909 (2013).
23. Hu, D. L. & Bush, J. W. M. Meniscus-climbing insects. *Nature* **437**, 733–736 (2005).
24. Brackenbury, J. Caterpillar kinematics. *Nature* **390**, 453 (1997).
25. Wang, W. *et al.* Locomotion of inchworm-inspired robot made of smart soft composite (SSC). *Bioinspir. Biomim.* **9**, 046006 (2014).
26. Taylor, G. Analysis of the swimming of microscopic organisms. *Proc. R. Soc. Lond. Ser. A* **209**, 447–461 (1951).
27. Campbell, J. F. & Kaya, H. K. How and why a parasitic nematode jumps. *Nature* **397**, 485–486 (1999).
28. Zhang, X. D., Mehrtash, M. & Khamesee, M. B. Dual-axial motion control of a magnetic levitation system using Hall-effect sensors. *IEEE/ASME Trans. Mechatron.* **21**, 1129–1139 (2016).
29. Abbott, J. J. *et al.* How should microrobots swim? *Int. J. Robot. Res.* **28**, 1434–1447 (2009).
30. Aguilar, J. & Goldman, D. I. Robophysical study of jumping dynamics on granular media. *Nat. Phys.* **12**, 278–283 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements W.H. thanks the Alexander von Humboldt Foundation for financial support. This work is funded by the Max Planck Society. We thank Z. Burghard and A. Diem from the University of Stuttgart for evaluating the Young's modulus of our robots, K. Suppelt and S. Meyer from Fujifilm Visualsonics for their help with the ultrasound-guided experiments, and the members from Physical Intelligence Department at the Max Planck Institute for Intelligent Systems for their comments.

Author Contributions M.S., W.H., G.Z.L. and M.M. proposed and designed the research. W.H. performed all experiments. G.Z.L. developed all theoretical and empirical models, except for the meniscus-climbing model, which was developed by M.M. The experimental data were analysed by W.H., G.Z.L. and M.M. All authors wrote the paper and participated in discussions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.S. (sitti@is.mpg.de).

Reviewer Information *Nature* thanks K.-J. Cho, R. Kramer-Bottiglio and B. Mazzolai for their contribution to the peer review of this work.

Generating carbyne equivalents with photoredox catalysis

Zhaofeng Wang¹, Ana G. Herrai¹, Ana M. del Hoyo¹ & Marcos G. Suero¹

Carbon has the unique ability to bind four atoms and form stable tetravalent structures that are prevalent in nature. The lack of one or two valences leads to a set of species—carbocations, carbanions, radicals and carbenes—that is fundamental to our understanding of chemical reactivity¹. In contrast, the carbyne—a monovalent carbon with three non-bonded electrons—is a relatively unexplored reactive intermediate^{2–6}; the design of reactions involving a carbyne is limited by challenges associated with controlling its extreme reactivity and the lack of efficient sources^{7–9}. Given the innate ability of carbynes to form three new covalent bonds sequentially, we anticipated that a catalytic method of generating carbynes or related stabilized species would allow what we term an ‘assembly point’ disconnection approach for the construction of chiral centres. Here we describe a catalytic strategy that generates diazomethyl radicals as direct equivalents of carbyne species using visible-light photoredox catalysis. The ability of these carbyne equivalents to induce site-selective carbon–hydrogen bond cleavage in aromatic rings enables a useful diazomethylation reaction, which underpins sequencing control for the late-stage assembly-point functionalization of medically relevant agents. Our strategy provides an efficient route to libraries of potentially bioactive molecules through the installation of tailored chiral centres at carbon–hydrogen bonds, while complementing current translational late-stage functionalization processes¹⁰. Furthermore, we exploit the dual radical and carbene character of the generated carbyne equivalent in the direct transformation of abundant chemical feedstocks into valuable chiral molecules.

Given that the simplest methylidyne ($\text{:}\dot{\text{C}}\text{--H}$) was one of the first molecules detected in interstellar space^{11,12}, and is considered one of the basic ingredients of life^{13,14}, it is surprising that carbynes have remained largely unexplored in the design and invention of reactions for over fifty years. Pioneering efforts by Strausz and Patrick in the 1960s and 1970s for the generation of a doublet ground-state carbyne ($\text{:}\dot{\text{C}}\text{--CO}_2\text{Et}$), using $\text{Hg}\{\text{C(=N}_2\text{)CO}_2\text{Et}\}_2$ and ultraviolet light, established that these species had a dual radical and carbene behaviour^{2–6}. The monovalent carbyne species promise to unveil new chemical reactivity at a single carbon atom, complementing the reactivity of traditional divalent and trivalent reactive carbon species (Fig. 1a). However, the absence of efficient carbyne sources and generation strategies preclude this development.

We were drawn to three fundamental features of carbyne intermediates: (i) their highly electrophilic nature; (ii) their distinct radical and carbene reactivities on the same carbon atom; and (iii) their natural ability to form three sigma bonds to complete the valence shell. We wondered whether the construction of chiral centres using aromatic chemical feedstocks could be achieved by using the assembly-point functionalization of carbynes (Fig. 1b). Drawing inspiration from Strausz², we speculated that the monovalent carbon could be decorated with two orthogonal leaving groups that, upon subsequent catalytic activations, would reveal the dual radical/carbene reactivity and underpin sequence control for a desired chiral centre (Fig. 1c). We initially

hypothesized that a diazo compound bearing an appropriate redox-active leaving group could generate—through catalytic single-electron reduction—a diazomethyl radical, which could induce aryl C–H bond cleavage. After this, the prochirality of the diazo-functionalized intermediate would be exploited using the broad range of known metal-catalysed processes¹⁵.

One of the most important features of the methodology of photoredox catalysis^{16,17} is the ability to generate radical cation/anion species via single-electron transfer processes and transient radical species under mild conditions. In this context, hypervalent iodine reagents¹⁸ have shown broad application as a source of carbon-centred radical species owing to their unique combination of redox properties, stability and availability¹⁹. From appropriate hypervalent iodine precursors, we prepared new bench-stable benziodoxolone **1a** and its pseudocyclic analogue **1b**^{20,21}, and hypothesized that they would be able to generate the envisaged substituted diazomethyl radical $\text{N}_2 = \text{C}(\bullet)\text{--CO}_2\text{Et}$ by catalytic photoredox activation (Fig. 2a). A technique often used to determine the feasibility of the latter activation is based on fluorescent quenching studies (Stern–Volmer)¹⁷ of a photoredox complex ML_n , where M is a metal and L is a ligand. These studies evaluate the decrease of the intensity of the emission of the corresponding active photoexcited state $^*[\text{ML}_n]$ in the presence of a quencher. Initially, we performed Stern–Volmer studies with the commercial photocatalyst $\text{Ru}(\text{bpy})_3(\text{PF}_6)_2$ (where bpy is 2,2′-bipyridine) and no fluorescence quenching was observed for the active photoexcited state $^*[\text{Ru}(\text{bpy})_3]^{2+}$ (maximum emission wavelength 615 nm) with **1a**. We reasoned that a single-electron transfer event might be thermodynamically uphill, on the basis of the redox potential of $[\text{Ru}(\text{bpy})_3]^{2+}$ (half-wave potential $E_{1/2}^{(\text{II})/(\text{III})} = -0.81$ V versus saturated calomel electrode (SCE), CH_3CN) and reagent **1a** (reduction potential $E_{\text{red}} = -0.95$ V versus SCE, CH_3CN). However, with 1 equivalent (equiv.) of $\text{Zn}(\text{NTf}_2)_2$ (activator of cyclic hypervalent reagents, where Tf is trifluoromethanesulfonyl) with **1a** or the more oxidative reagent **1b** ($E_{\text{red}} = -0.33$ V versus SCE, CH_3CN), we found clear deactivation of $^*[\text{Ru}(\text{bpy})_3]^{2+}$ (see Supplementary Information for details). We later provided evidence of the formation of the envisaged radical $\text{N}_2 = \text{C}(\bullet)\text{--CO}_2\text{Et}$ (intermediate **int-I**) by using 1 equiv. of $\text{Ru}(\text{bpy})_3(\text{PF}_6)_2$ and **1a** or **1b** under white light-emitting diode (LED) irradiation (Fig. 2b). The formation of diethyl acetylenedicarboxylate was rationalized through a sequential radical homodimerization of **int-I** and dinitrogen elimination in bis-diazo **int-II**.

After this, we envisioned the following hypothesis for an aromatic C–H bond diazomethylation reaction (Fig. 2c). Irradiation of $[\text{Ru}(\text{bpy})_3]^{2+}$ (**2**) with visible light would generate $^*[\text{Ru}(\text{bpy})_3]^{2+}$ (**3**), which would undergo single-electron reduction with reagents **1** and generate a transient anion radical evolving to iodoarene **4** and **int-I**. The latter intermediate would intercept an aromatic ring (**6**) and generate the cyclohexadienyl radical **int-III**. Subsequently, oxidation with $[\text{Ru}(\text{bpy})_3]^{3+}$ (**5**, $E_{1/2}^{(\text{III})/(\text{II})} = 1.29$ V versus SCE in CH_3CN) and proton elimination would lead to the expected diazo compound **7** and the $\text{Ru}(\text{II})$ catalyst **2**.

¹Institute of Chemical Research of Catalonia (ICIQ), The Barcelona Institute of Science and Technology, Avinyuda Països Catalans 16, 43007 Tarragona, Spain.

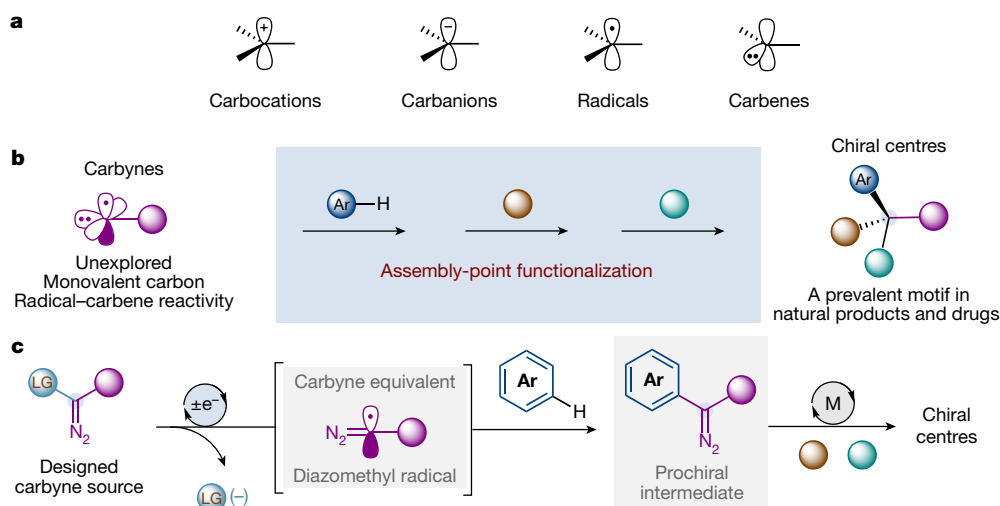


Figure 1 | Monovalent carbyne species enable an assembly-point functionalization strategy for chiral centre construction with aryl C-H bonds. **a**, Fundamental carbon-based reactive di- and tri-valent species. **b**, New disconnection approach for chiral centre construction

with monovalent carbyne species. **c**, Designed carbyne source enables sequencing control strategy using orthogonal catalytic activation modes. **Ar**, arene; LG, leaving group; M, metal catalyst.

Our envisaged C-H diazomethylation was first evaluated using the chemical feedstock 1,4-dimethylbenzene (*p*-xylene) and reagents **1a** and **1b**. We identified the best results by using the combination of **1b** (1 equiv.), *p*-xylene (2 equiv.), sodium bicarbonate as base (3 equiv.), CH_3CN (0.1 M) and $\text{Ru}(\text{bpy})_3(\text{PF}_6)_2$ (1 mol%) under white LED irradiation. The product **7a** was obtained in 73% isolated yield after purification with column chromatography (Fig. 3a).

Other Ru photocatalysts [$\text{Ru}(\text{dtbbpy})_3(\text{PF}_6)_2$, 63% yield, where dtbbpy is 4,4'-di-*tert*-butyl-2,2'-bipyridine; $\text{Ru}(\text{dMebpy})_3(\text{PF}_6)_2$, 68% yield, where dMebpy is 4,4'-dimethyl-2,2'-bipyridyl] organic bases (2,6-di-*tert*-butylpyridine, 63% yield), water as co-solvent (dichloromethane/ H_2O , 52% yield) or linear hypervalent iodine reagents (**1j-k**, 41%–54% yield) were also able to provide compound **7a** as a single reaction product. Full details of the optimization are provided in Supplementary Tables 1–3.

In addition, whereas the increased concentration of *p*-xylene in the reaction (10 equiv.) showed a dramatic improvement of the efficiency

of the process (96% yield), only a moderate yield of 54% was obtained when *p*-xylene (1 equiv.) was the limiting reagent. Having the optimized reaction conditions in hand, we next evaluated the scope of the photocatalytic C-H diazomethylation reaction. As shown in Fig. 3a, this process was successfully applied in a wide range of aromatic hydrocarbons decorated with a variety of useful functional groups. With di-substituted arenes (**7b-i**), we observed a high preference of the radical species $\text{N}_2=\text{C}(\bullet)-\text{CO}_2\text{Et}$ (**int-I**, Fig. 2) to react at electron-rich sites; a reasonable behaviour considering the electrophilic nature of this carbyne equivalent. Remarkably, sterically congested substrates produced the corresponding diazo compounds with a high degree of efficiency (**7j**, 99% yield).

A substantial selectivity challenge was presented in the C-H functionalization of mono-substituted arenes, which generally lead to isomeric mixtures in electrophilic radical substitutions. However, we were pleased to find that the C-H diazomethylation reaction generally occurred with high *ortho*-selectivity in arenes substituted with

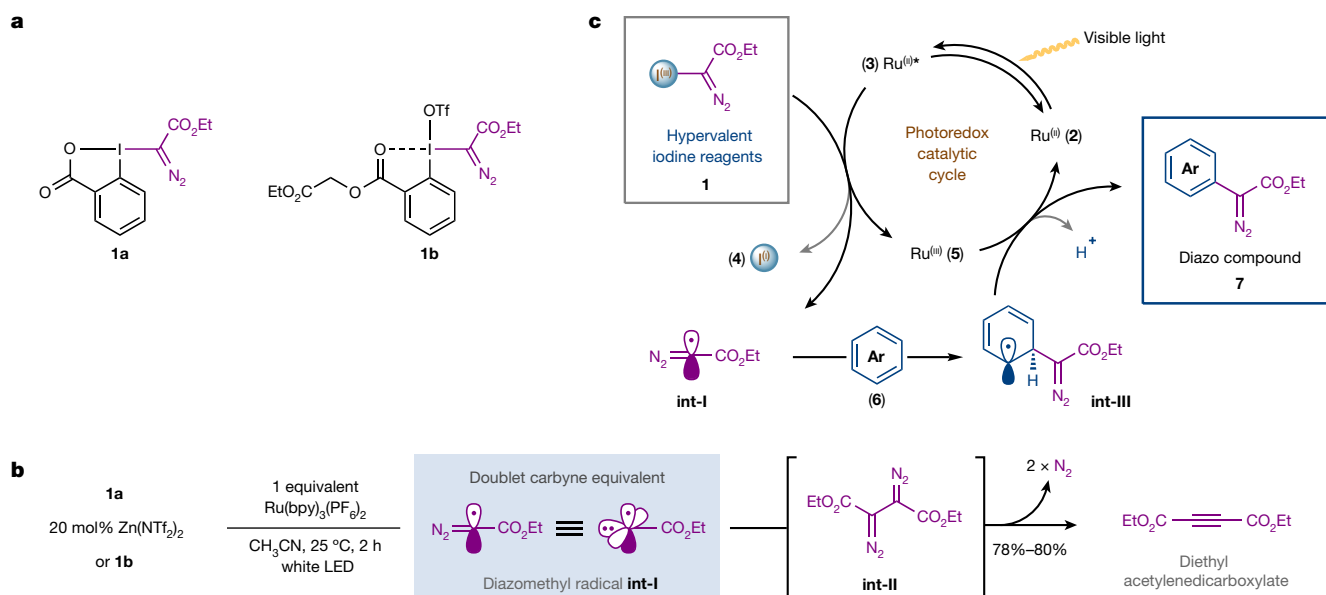


Figure 2 | New hypervalent iodine reagents and photoredox catalysis enables a C-H bond diazomethylation reaction. **a**, Hypervalent iodine reagents **1a** and **1b** as carbyne sources. These diazo transfer reagents are

easy-to-prepare solids and stable at 20 °C. **b**, Photoredox generation of diazomethyl radical **int-I**. **c**, Working mechanistic proposal.

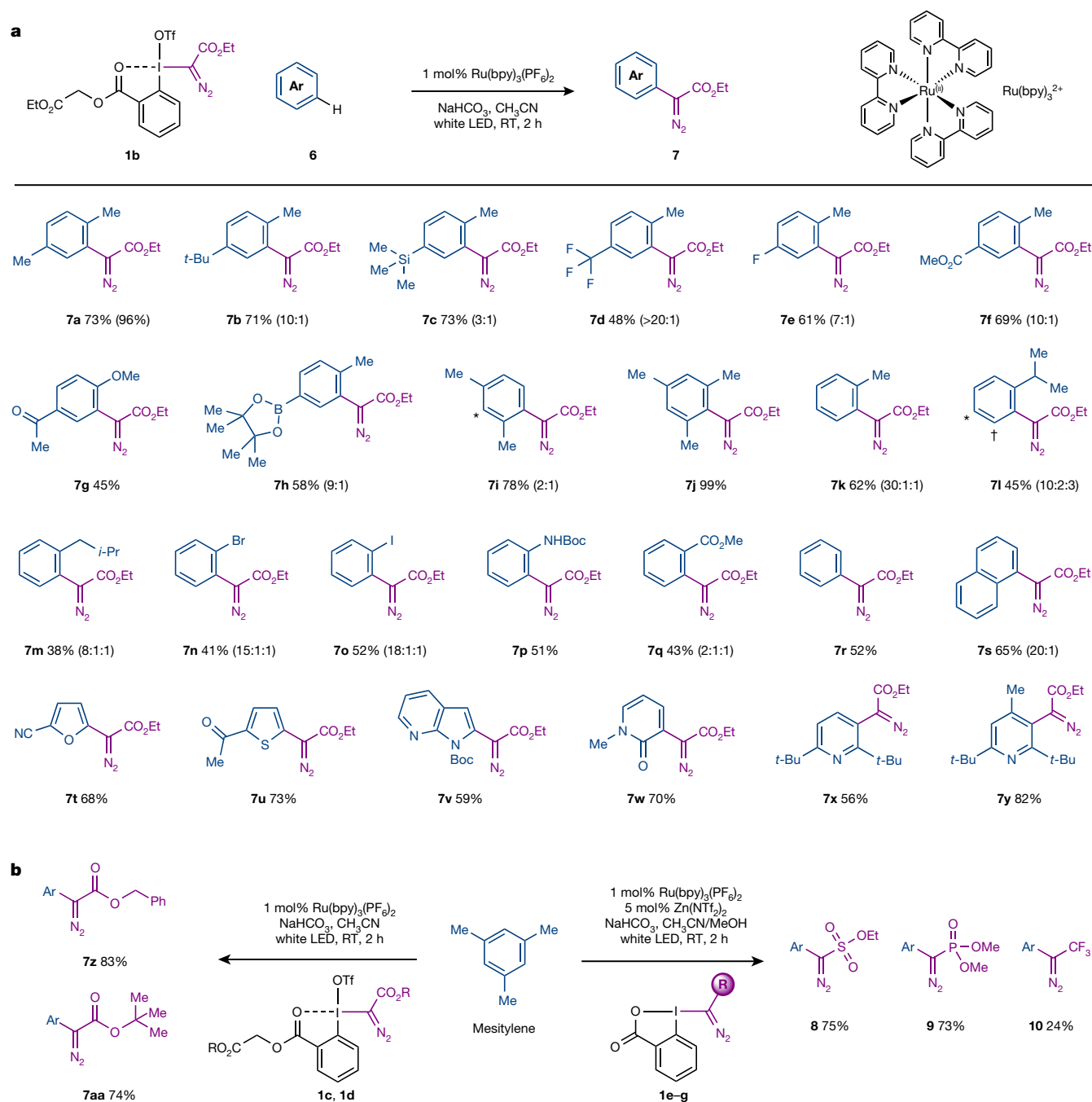


Figure 3 | Arene C–H diazomethylation by means of photoredox catalysis. **a**, Scope of arene **6**. In **7a–c**, **7g–m**, **7o**, **7p**, and **7s**, 2 equiv. of arene were used. In **7a**, the yield in parentheses was obtained with 10 equiv. of *p*-xylene. In **7d–f**, **7n**, **7q** and **7r**, 10 equiv. of arene were used. In **7t**, **7u** and **7w**, 5 equiv. of heteroarene were used. In **7x** and **7y**,

3 equiv. of heteroarene were used. RT, room temperature. In **7s**, 2,6-di-*tert*-butylpyridine was used as the base. In **7w**, Na_2CO_3 was used as the base. In **7l** and **7i**, the dagger and asterisk indicate the minor isomers. **b**, Scope of reagents **1**. Experiments were replicated at least twice for consistency.

alkyl groups (**7k–m**), halogens (**7n–o**), amides (**7p**) or carbonyls (**7q**). The selectivity observed here is in sharp contrast with the *para*-selective functionalization of aromatic rings with high-electron-affinity radicals²². Additionally, non-substituted arenes such as benzene (**7r**) or naphthalene (**7s**) and five- and six-membered heterocycles (**7t–y**) also worked well.

We then explored the scope of the hypervalent iodine reagents **1c–g** by using mesitylene as the arene substrate (Fig. 3b). We observed that pseudocyclic reagent analogues **1c** ($R = CH_2Ph$) and **1d** ($R = CMe_3$) were effective for transferring alternative ester functionalities (**7z–aa**; 83%–74% yield, respectively). Moreover, cyclic reagents **1e–g** permitted the synthesis of useful diazo compounds substituted with sulfonates

(**8**), phosphonates (**9**) or trifluoromethyl groups (**10**) via new carbyne equivalents $N_2 = C(\bullet) - R$ ($R = SO_2(OEt)$, $PO(OMe)_2$, CF_3) (see Fig. 3b). Our methodology for the synthesis of valuable diazo compounds complements current strategies involving palladium-catalysed cross-couplings with aryl iodides and diazoacetates^{23,24} or diazo-transfer processes with aryl acetates and sulfonyl azides¹⁵, but it also provides important synthetic advances because: (i) it precludes the need of pre-functionalized aromatic starting materials (aryl iodides or aryl acetates), (ii) it enables access to *ortho*-substituted and sterically congested aryl diazo compounds (elusive via cross-coupling reactions) and (iii) it streamlines the synthesis of the less-accessible diazo compounds **8–10** in only one synthetic step (elusive via cross-coupling reactions).

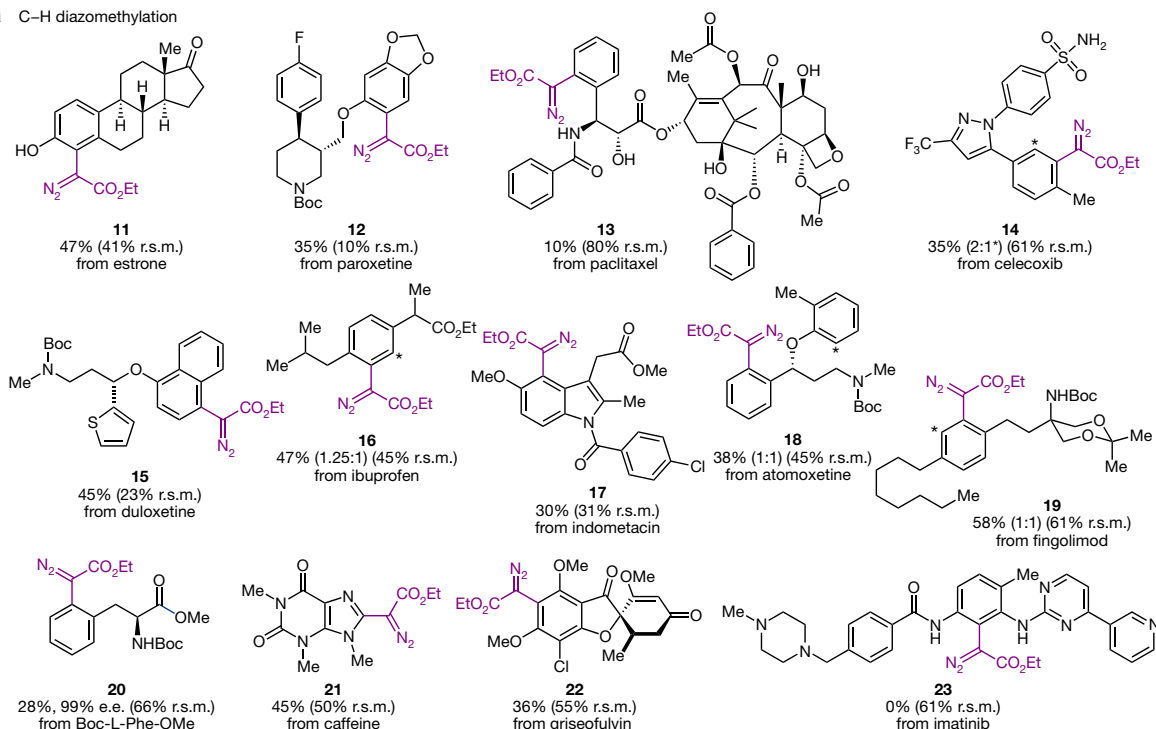
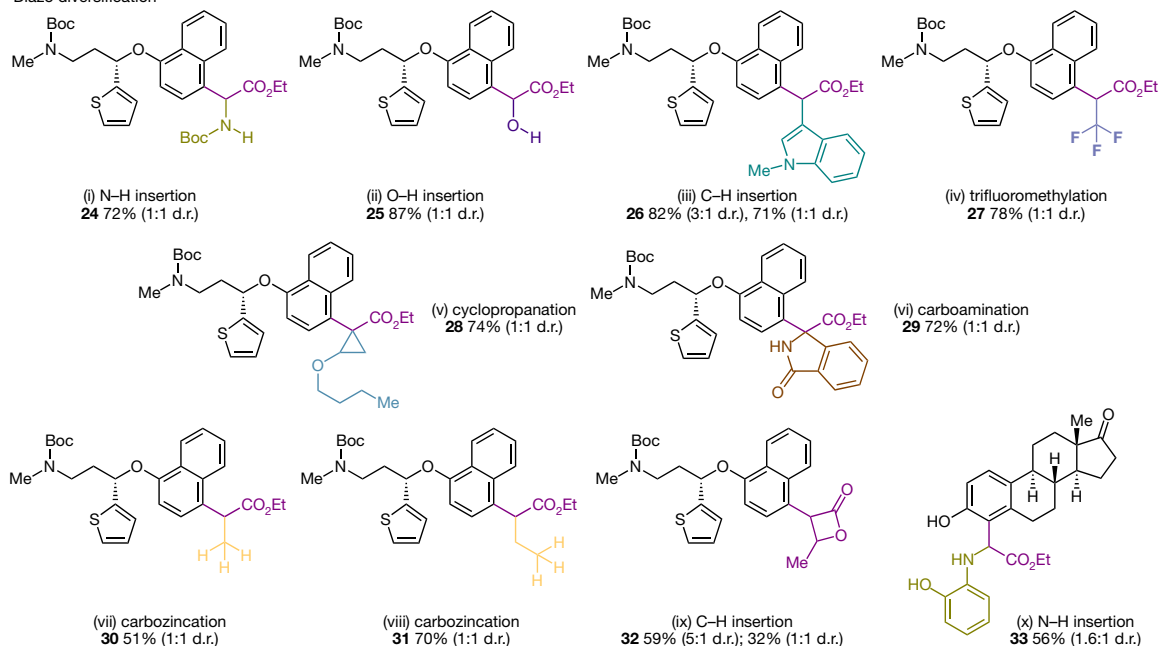
a C–H diazomethylation**b** Diazo diversification

Figure 4 | Late-stage assembly-point diversification of medically relevant agents. a, Late-stage C–H diazomethylation. r.s.m., recovered starting material. Reaction conditions: Ru(bpy)₃(PF₆)₂ (0.1 mol%), medically relevant agent (1 equiv.), **1b** (2 equiv.), NaHCO₃ (3 equiv.), CH₃CN, RT, 1–2 h. In **16–20** and **22**, 2 equiv. of medically relevant agent and 1 equiv. of **1b** were used. In **14**, isomers were separated by semi-preparative HPLC (high-performance liquid chromatography). In **16**, isomers were separated by column chromatography. In **19**, isomers were separated by preparative thin layer chromatography. In **14**, **16**, **18** and **19**, the asterisk indicates another isomer obtained. **b**, Diversification of **15** and **11**. The coloured components of each compound indicate the new fragments that have been added. Reaction conditions are as follows. (i) [RuCl₂(*p*-cymene)]₂ (1 mol%), *t*-BuOCONH₂ (1.5 equiv.), dichloromethane, RT, 3 h. (ii) Rh₂(OAc)₄ (1 mol%), water (3 equiv.), –50 °C to RT, 2 h. (iii) ³Rh₂(OAc)₄ (1 mol%), (*R*)-3,3'-bis(2,4,6-trisopropylphenyl)-1,1'-binaphthyl-2,2'-diylhydrogenphosphate [(*R*)-TRIP] (2 mol%), 1-methylindole (1.3 equiv.), 4 Å molecular sieves, toluene, RT, 1.5 h; ³Rh₂(OAc)₄ (1 mol%),

1-methylindole (1.3 equiv.), 4 Å molecular sieves, toluene, RT, 6 h. (iv) CuI (1.5 equiv.), Me₃SiCF₃ (1.65 equiv.), CsF (1.65 equiv.), *N*-methyl-pyrrolidine, RT, 30 min, then addition of **15** (1 equiv.), water (44 equiv.), RT, 20 h. (v) Rh₂(O₂CCF₃)₄ (1 mol%), butyl vinyl ether (5 equiv.), dichloromethane, RT, 20 h. (vi) [RhCp*Cl₂]₂ (1 mol%), Cs₂CO₃ (20 mol%), *N*-(pivaloyloxy) benzamide (1 equiv.), MeCN, 70 °C, 3 h. (vii) Rh₂(TPA)₄ (1 mol%), ZnMe₂ (2.4 equiv.), toluene, RT, 3.5 h. (viii) Rh₂(TPA)₄ (1 mol%), ZnEt₂ (2.4 equiv.), toluene, RT, 3.5 h. (ix) With tetrakis[(*S*)-(+)-[(1*S*)-1-(4-bromophenyl)-2,2-diphenylcyclopropanecarboxylate]dirhodium(II) Rh₂(S-BTPCP)₄ (1 mol%), dichloromethane, RT, 3 h, the yield was 59%; with bis[rhodium(α,α,α', α'-tetramethyl-1,3-benzenedipropionic acid)] Rh₂(esp)₂ (1 mol%) dichloromethane, RT, 3 h, the yield was 32%. (x) Cu(MeCN)₄PF₆ (1 mol%), 2-aminophenol (1.2 equiv.), MeOH, 0 °C to RT, 2 h. In **24**, isomers were separated by semi-preparative HPLC. d.r., diastereomeric ratio. Enantiomeric excess (e.e.) was determined by HPLC analysis on a chiral stationary phase. Cp*, pentamethylcyclopentadienyl.

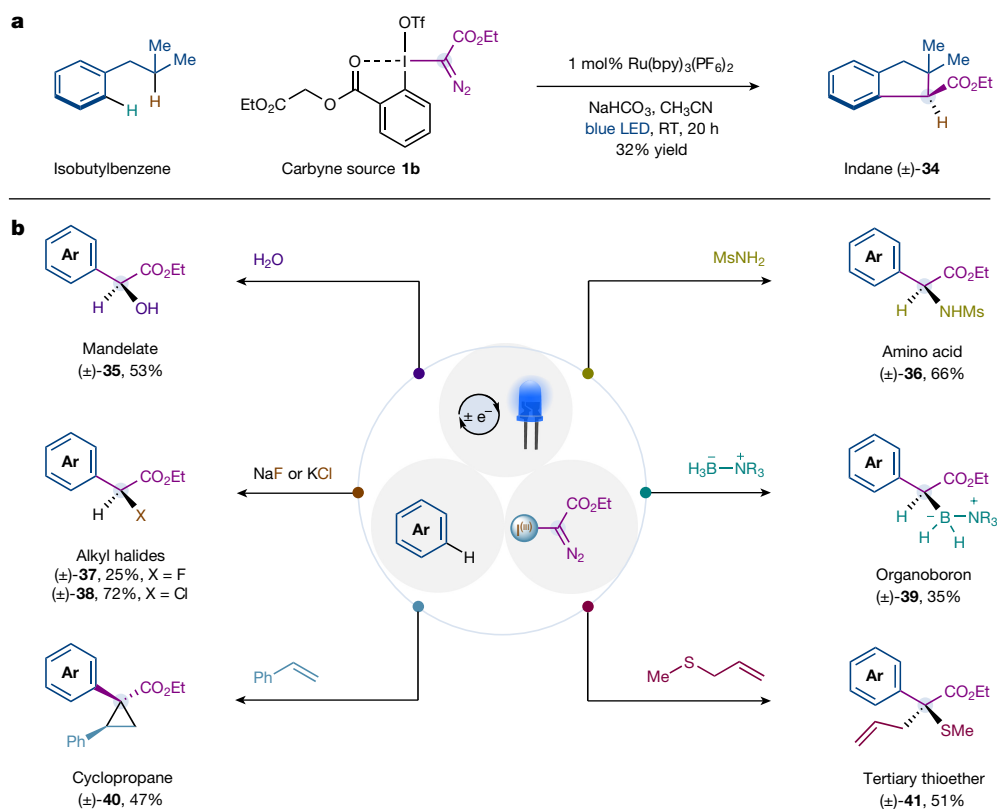


Figure 5 | Catalytic assembly-point functionalization of carbyne equivalents with feedstock chemicals. **a**, Synthesis of indane (±)-**34** by double C–H bond functionalization. **b**, *p*-xylene, reagent **1b** and the reaction conditions shown in **a** enabled the synthesis of chiral derivatives

(±)-**35–41** (see Supplementary Information for details). *p*-xylene was used as the aromatic hydrocarbon. The aryl substituent in **35–41** corresponds to 2,5-dimethylphenyl. NR₃ = *N*-methylpyrrolidine.

At present, more than two-thirds of prescription drugs contain at least one chiral centre, and their presence in drug candidates is linked with the success of transition from the discovery stage to clinical testing²⁵. Our assembly-point functionalization strategy (Fig. 1b), based on the C–H diazomethylation reaction developed, now opens up an opportunity to introduce chiral centres in aromatic rings and generate new analogues of a lead structure without resorting to *de novo* synthesis. We evaluated a selection of medically relevant agents to demonstrate its potential amenability for late-stage functionalization applications (Fig. 4a). Subjection of a range of drugs, including paclitaxel (antitumoral), duloxetine (antidepressant) or fingolimod (treatment of multiple sclerosis), to our standard protocol with **1b** resulted in the successful formation of the corresponding diazomethylated drug molecules and natural products (**11–22**, 10%–58% yield).

The excellent selectivity profile observed may be governed by the innate radicophilicity of the electron-rich aromatics, promoting reaction with the electrophilic radical N₂ = C(•)–CO₂Et (**int-I**). Furthermore, the mono C–H diazomethylation achieved is striking considering the presence of multiple aromatic rings in some of the abovementioned drugs (**11–22**), but also their broad compatibility with functional groups. Although conversions are in general modest, or low in some cases, the starting drug molecules can usually be recovered and recycled from the reaction crude by column chromatography. Moreover, *de novo* synthesis of diazo compounds **11–22** would not only be tedious but may also be challenging considering that the diazo group may not be compatible with the corresponding synthetic sequence. In contrast, drug molecules bearing basic nitrogen functionalities such as imatinib were not suitable for this process (**23**, Fig. 4a, see also Supplementary Fig. 3). Next, we transformed the diazo functionality installed in **15** into a suite of diverse chiral centres decorated with useful pharmacophores by exploiting a range of robust catalytic N–H (**24**, 72% yield), O–H (**25**, 87% yield) and C–H insertion reactions

(**26**, 82% yield), as well as trifluoromethylation (**27**, 78% yield), cyclopropanation (**28**, 74% yield) and carboamination (**29**, 72% yield) (Fig. 4b).

The introduction of a simple methyl group into lead compounds in drug discovery programmes has been shown to boost binding potency and lower the half-maximal inhibitory concentration (IC₅₀ value); this is the so-called “methyl magic effect”²⁶. The decoration of **15** with a methyl group would enhance the value and broad utility of our assembly-point strategy for library synthesis and complement late-stage alkylation strategies²⁷. However, although a direct metal-carbene C–H insertion reaction with CH₄ would introduce a methyl group in **15**, only ethyl diazoacetate (N₂ = C(H)–CO₂Et) has been reported to work in combination with silver catalysis under high methane/CO₂ pressures²⁸. In the search for a simple solution, we found that a commercial ZnMe₂ or ZnEt₂ and Rh₂(TPA)₄ catalyst (where TPA is triphenylacetate) allowed the successful introduction of a methyl or ethyl group at room conditions (**30**, 51% yield; **31**, 70% yield) (Fig. 4b)²⁹. The application of diorganozinc reagents in the functionalization of complex molecules has recently been demonstrated³⁰.

In addition, an intramolecular Rh-catalysed C–H insertion reaction with **15** allowed access to a privileged oxetane core and the Cu-catalysed N–H insertion in **11** demonstrated that one can selectively perform metal-catalysed carbene insertion reactions in complex molecules. The diazo functionalizations presented in Fig. 4b streamline the access to two novel diastereoisomers of duloxetine derivatives, which may be appealing from a medicinal chemistry point of view. We also showed that chiral catalysts are able to override the chiral information of **15** and favour one isomer (**26**, 82% yield, 3:1 diastereomeric ratio d.r.; **32**, 59% yield, 5:1 d.r.).

Finally, we questioned whether we could directly convert abundant aromatic hydrocarbons into valuable chiral molecules by using the unexplored dual radical and carbene character of N₂ = C(•)–CO₂Et

(**int-I**). To this end, isobutylbenzene was selectively converted to indane (\pm)-**34** (32% yield), a core that is prevalent in natural products, medicines and materials (Fig. 5a). This process occurred via double site-selective C–H functionalization and represents a unique example of a carbyne pentannulation. In this scenario, we successfully found that by using the same reaction conditions, *p*-xylene, **1b** and a range of simple reagents (water, methanesulfonamide, NaF, KCl, styrene, borane *N*-methylpyrrolidine complex or allyl methyl sulphide) led to a range of valuable chiral building blocks (\pm)-**35–41** (Fig. 5b).

These results clearly show that we can efficiently exert sequencing control in an assembly-point functionalization process of carbyne equivalents by means of photoredox catalysis. Key to our success was the use of blue LEDs as a source of visible light, which decomposes the diazo group into a free carbene in the corresponding aryl diazo intermediate (control experiments are provided in Supplementary Information)³¹. The latter processes clearly show the distinct dual radical–carbene reactivity of the monovalent carbon equivalent and its application in discovering previously elusive disconnection approaches.

Data Availability The data supporting the findings of this study are available within the paper and its Supplementary Information. Metrical parameters for the structure of reagents **1a**, **1b**, **1d**, **1f**, **1g** and products **7g** and **7u** (see Supplementary Information) are available free of charge from the Cambridge Crystallographic Data Centre (<https://www.ccdc.cam.ac.uk/>) under reference numbers CCDC 1548798, CCDC 1548799, CCDC 1566194, CCDC 1548843, CCDC 1548800, CCDC 1548797 and CCDC 1566193, respectively.

Received 6 June; accepted 22 November 2017.

- Trost, B. M. & Fleming, I. *Comprehensive Organic Synthesis* (Pergamon, 1991).
- Thap, D. M., Gunning, H. E. & Strausz, O. P. Formation and reactions of monovalent carbon intermediates. I. Photolysis of diethyl mercuribisdiazoacetate. *J. Am. Chem. Soc.* **89**, 6785–6787 (1967).
- Strausz, O. P., Thap, D. M. & Font, J. Formation and reactions of monovalent carbon intermediates. II. Further studies on the decomposition of diethyl mercuribisdiazoacetate. *J. Am. Chem. Soc.* **90**, 1930–1931 (1968).
- Strausz, O. P. *et al.* Formation and reactions of monovalent carbon intermediates. III. Reaction of carboxymethylene with olefins. *J. Am. Chem. Soc.* **96**, 5723–5732 (1974).
- Patrick, T. B. & Kovitch, G. H. Photolysis of diethyl mercuribisdiazoacetate and ethyl diazoacetate in chloroalkanes. *J. Org. Chem.* **40**, 1527–1528 (1975).
- Patrick, T. B. & Wu, T.-T. Photodecomposition of diethyl mercuribis(diazoacetate) in several heterocyclic systems. *J. Org. Chem.* **43**, 1506–1509 (1978).
- Fürstner, A. Alkyne metathesis on the rise. *Angew. Chem. Int. Ed.* **52**, 2794–2819 (2013).
- Bino, A., Ardon, M. & Shirman, E. Formation of a carbon-carbon triple bond by coupling reactions in aqueous solution. *Science* **308**, 234–235 (2005).
- Bogoslavsky, B. *et al.* Do carbyne radicals really exist in aqueous solution? *Angew. Chem. Int. Ed.* **51**, 90–94 (2012).
- Cernak, T., Dykstra, K. D., Tyagarajan, S., Vachal, P. & Krska, S. W. The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* **45**, 546–576 (2016).
- Swings, P. & Rosenfeld, L. Considerations regarding interstellar molecules. *Astrophys. J.* **86**, 483–486 (1937).
- Rydbeck, O. E. H., Elder, J. & Irvine, W. M. Radio detection of interstellar CH. *Nature* **246**, 466–468 (1973).
- Morris, P. W. *et al.* Herschel/HIFI spectral mapping of C⁺, CH⁺, and CH in Orion BN/KL: the prevailing role of ultraviolet irradiation in CH⁺ formation. *Astrophys. J.* **829**, 15–46 (2016).
- Landau, E. Building Blocks of Life's Building Blocks Come From Starlight. <https://www.jpl.nasa.gov/news/news.php?feature=6645> (JPL, 2016).
- Ford, A. *et al.* Modern organic synthesis with α -diazocarbonyl compounds. *Chem. Rev.* **115**, 9981–10080 (2015).
- Nicewicz, D. A. & MacMillan, D. W. C. Merging photoredox catalysis with organocatalysis: the direct asymmetric alkylation of aldehydes. *Science* **322**, 77–80 (2008).
- Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113**, 5322–5363 (2013).
- Yoshimura, A. & Zhdankin, V. V. Advances in synthetic applications of hypervalent iodine compounds. *Chem. Rev.* **116**, 3328–3435 (2016).
- Wang, L. & Liu, J. Synthetic applications of hypervalent iodine(III) reagents enabled by visible light photoredox catalysis. *Eur. J. Org. Chem.* **2016**, 1813–1824 (2016).
- Weiss, R., Seubert, J. & Hampel, F. α -Aryliodonio diazo compounds: S_N reactions at the α -C atom as a novel reaction type for diazo compounds. *Angew. Chem. Int. Ed. Engl.* **33**, 1952–1953 (1994).
- Schnaars, C., Hennum, M. & Bonge-Hansen, T. Nucleophilic halogenations of diazo compounds, a complementary principle for the synthesis of halodiazo compounds: experimental and theoretical studies. *J. Org. Chem.* **78**, 7488–7497 (2013).
- Boursalian, G. B., Ham, W. S., Mazzotti, A. R. & Ritter, T. Charge-transfer-directed radical substitution enables para-selective C–H functionalization. *Nat. Chem.* **8**, 810–815 (2016).
- Ye, F. *et al.* Palladium-catalyzed C–H functionalization of acyldiazomethane and tandem cross-coupling reactions. *J. Am. Chem. Soc.* **137**, 4435–4444 (2015).
- Fu, L., Mighion, J. D., Voight, E. A. & Davies, H. M. L. Synthesis of 2,2,2-trichloroethyl aryl- and vinyl diazoacetates by palladium-catalyzed cross-coupling. *Chem. Eur. J.* **23**, 3272–3275 (2017).
- Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52**, 6752–6756 (2009).
- Schönherr, H. & Cernak, T. Profound methyl effects in drug discovery and a call for new C–H methylation reactions. *Angew. Chem. Int. Ed.* **52**, 12256–12267 (2013).
- Le, C., Liang, Y., Evans, R. W., Li, X. & Macmillan, D. W. C. Selective sp³ C–H alkylation via polarity-match-based cross-coupling. *Nature* **547**, 79–83 (2017).
- Caballero, A. *et al.* Silver-catalyzed C–C bond formation between methane and ethyldiazoacetate in supercritical CO₂. *Science* **332**, 835–838 (2011).
- Panish, R., Selvaraj, R. & Fox, J. M. Rh(II)-catalyzed reactions of diazoesters with organozinc reagents. *Org. Lett.* **17**, 3978–3981 (2015).
- Qin, T. *et al.* A general alkyl-alkyl cross-coupling enabled by redox-active esters and alkylzinc reagents. *Science* **352**, 801–805 (2016).
- Zhu, Z., Bally, T., Stracener, L. & McMahon, R. J. Reversible interconversion between singlet and triplet 2-naphthyl(carbomethoxy)carbene. *J. Am. Chem. Soc.* **121**, 2863–2874 (1999).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by the ICIQ Foundation, the CERCA Programme (Generalitat de Catalunya), MINECO (CTQ2016-75311-P, AEI/FEDER-EU; Severo Ochoa Excellence Accreditation 2014–2018, SEV-2013-0319), the CELLEX Foundation through the CELLEX-ICIQ high-throughput experimentation platform. We thank the European Union for a Marie Curie-COFUND post-doctoral fellowship (to Z.W.) and the CELLEX Foundation for pre-doctoral (to A.G.H.) and post-doctoral fellowships (to A.M.d.H.). We thank the ICIQ Research Support Area, and F. Bravo for LC/MS instrumentation. M.G.S. is a Junior Group Leader of the ICIQ Starting Career Programme 2014–2019.

Author Contributions M.G.S. conceived the idea of developing new hypervalent iodine reagents for the generation of carbynes. Z.W., A.G.H. and A.M.d.H. performed the experiments. M.G.S. wrote the manuscript. All authors contributed to the analysis and interpretation of the data and commented on the final draft of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.G.S. (mgsuero@iciq.es).

Reviewer Information Nature thanks I. Larrosa and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Reconciling divergent trends and millennial variations in Holocene temperatures

Jeremiah Marsicek¹, Bryan N. Shuman¹, Patrick J. Bartlein², Sarah L. Shafer³ & Simon Brewer⁴

Cooling during most of the past two millennia has been widely recognized^{1,2} and has been inferred to be the dominant global temperature trend of the past 11,700 years (the Holocene epoch)³. However, long-term cooling has been difficult to reconcile with global forcing⁴, and climate models consistently simulate long-term warming⁴. The divergence between simulations and reconstructions emerges primarily for northern mid-latitudes, for which pronounced cooling has been inferred from marine and coastal records using multiple approaches³. Here we show that temperatures reconstructed from sub-fossil pollen from 642 sites across North America and Europe closely match simulations, and that long-term warming, not cooling, defined the Holocene until around 2,000 years ago. The reconstructions indicate that evidence of long-term cooling was limited to North Atlantic records. Early Holocene temperatures on the continents were more than two degrees Celsius below those of the past two millennia, consistent with the simulated effects of remnant ice sheets in the climate model Community Climate System Model 3 (CCSM3)⁵. CCSM3 simulates increases in ‘growing degree days’—a measure of the accumulated warmth above five degrees Celsius per year—of more than 300 kelvin days over the Holocene, consistent with inferences from the pollen data. It also simulates a decrease in mean summer temperatures of more than two degrees Celsius, which correlates with reconstructed marine trends and highlights the potential importance of the different subseasonal sensitivities of the records. Despite the differing trends, pollen- and marine-based reconstructions are correlated at millennial-to-centennial scales, probably in response to ice-sheet and meltwater dynamics, and to stochastic dynamics similar to the temperature variations produced by CCSM3. Although our results depend on a single source of palaeoclimatic data (pollen) and a single climate-model simulation, they reinforce the notion that climate models can adequately simulate climates for periods other than the present-day. They also demonstrate that amplified warming in recent decades increased temperatures above the mean of any century during the past 11,000 years.

Global cooling has previously been inferred for the Holocene, in a large part because of trends in a few northern mid-latitude temperature records³ (Fig. 1a–c); this cooling trend would not have been apparent without records with large cooling trends from the North Atlantic region^{3,6} (Fig. 1d). Furthermore, such cooling is inconsistent with the long-term warming that has been inferred from fossil pollen data across North America and Europe^{7,8}. There are numerous pollen records that span the continents, but only four were included in the global temperature reconstruction that was strongly influenced by North Atlantic cooling³. The independence of these pollen data from the marine record enables us to re-evaluate northern mid-latitude temperature trends. Independent marine and continental records also facilitate the detection of millennial-to-centennial-scale temperature variability, which remains poorly understood, but dominates global temperature reconstructions in the absence of records from the

North Atlantic³ (Fig. 1d). If cooling did not extend throughout the Holocene, then variability over timescales of more than 100 yr would be critical to differentiating the trend from the first 10,000 yr of the Holocene from that of the past 1,000 yr.

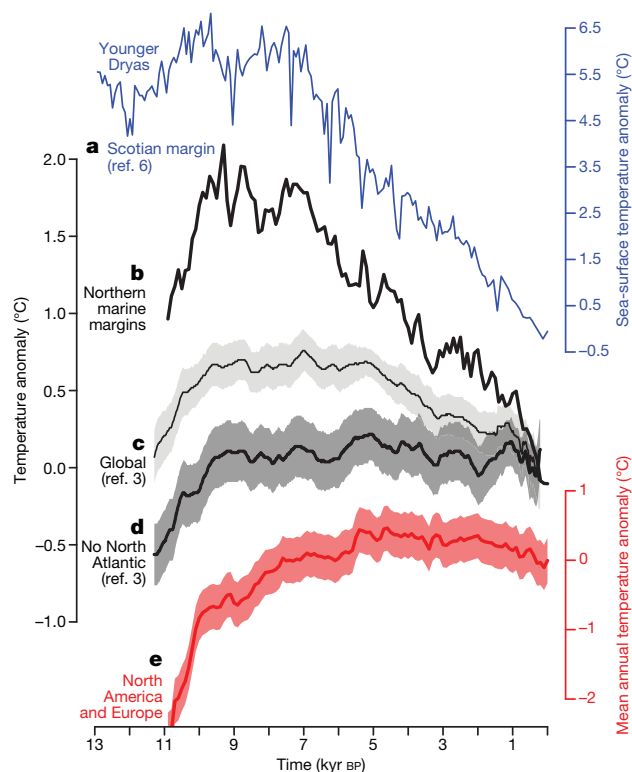


Figure 1 | Comparison of North Atlantic and global marine-margin temperature reconstructions with our pollen-inferred mean annual temperature reconstruction for North America and Europe. a–c, An example of a sea-surface temperature record from the North Atlantic (Scotian margin; a)⁶ containing a large secular cooling trend, which propagated into both the mean of northern marine-margin records (b) and a global mean temperature reconstruction³ (with 1σ uncertainty shaded; c). d, e, The global reconstruction from c, but with North Atlantic records removed³ (with 1σ uncertainty shaded; d) and our continental mean annual temperature reconstruction for North America and Europe (e; shading indicates 2.5%–97.5% uncertainty bands). The reconstruction in d shows no cooling trend, but displays multi-century variation including rapid warming before temperature maxima at around 9 kyr BP, 7 kyr BP and 5.5 kyr BP (these maxima are also evident in b and e). The reconstructions in b, d and e represent completely independent datasets. The Younger Dryas interval (indicated in a) is typically shown to have been colder than the past 500 years in the North Atlantic region. All anomalies are relative to the mean of AD 1450–1950.

¹Department of Geology and Geophysics, University of Wyoming, Laramie, Wyoming 82072, USA. ²Department of Geography, University of Oregon, Eugene, Oregon 97403, USA. ³US Geological Survey, Corvallis, Oregon 97331, USA. ⁴Department of Geography, University of Utah, Salt Lake City, Utah 84112, USA.

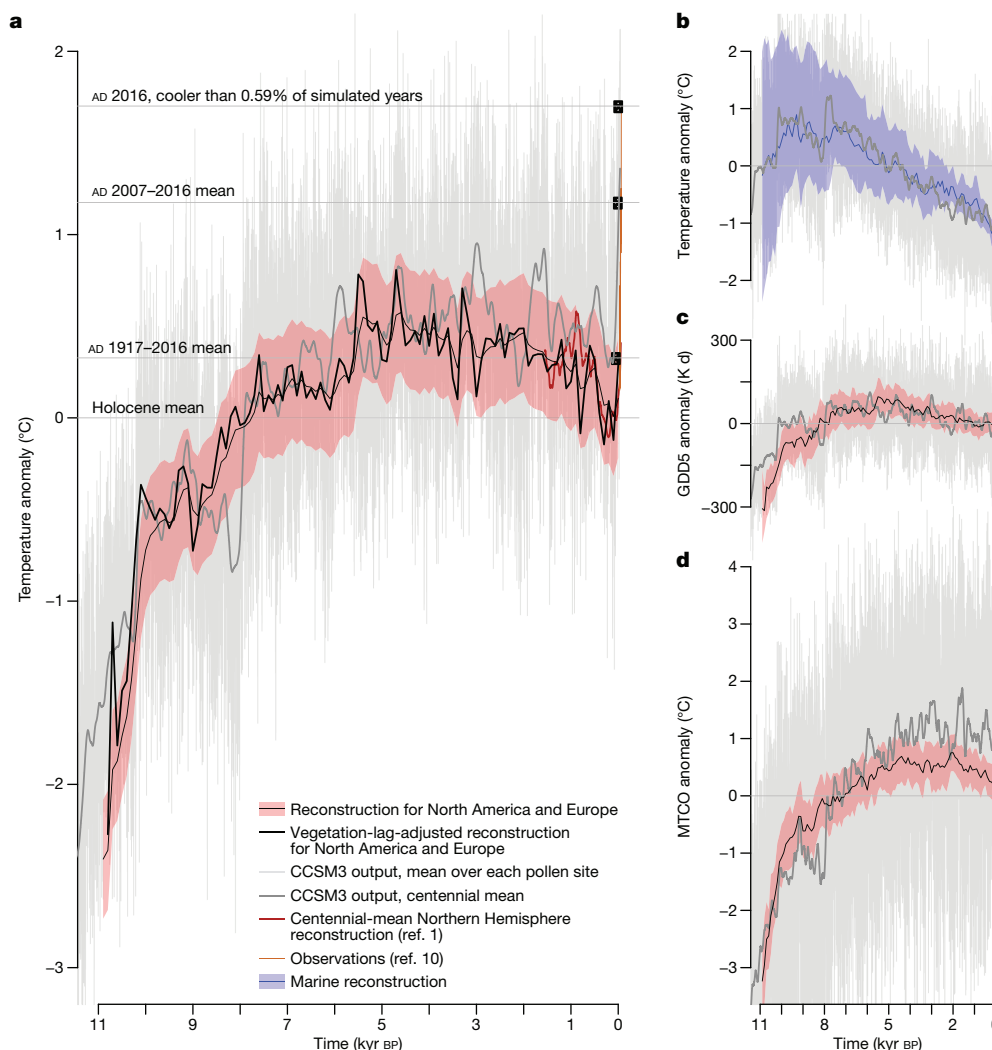


Figure 2 | Reconstructed and simulated annual and seasonal temperature changes in North America, Europe and the northern marine margins during the Holocene. a, c, d, Time series of pollen-inferred temperatures (thin black lines; red shading indicates the 2.5%–97.5% uncertainty bands) are compared to calendar-corrected CCSM3 output (thin grey lines indicate annual means; thick grey lines indicate 100-yr means) for mean annual temperature (a), growing degree days using a baseline of 5 °C (GDD5; c) and mean temperature of the coldest month (MTCO; d). **b,** The marine synthesis reconstruction (thin blue lines; blue shading indicates the 2.5%–97.5% uncertainty band derived from bootstrap resampling of the 20 marine datasets listed in Extended Data Table 2; also shown in Fig. 1b) is compared to the CCSM3

output for mean June–August land temperatures (grey lines as in a, c and d). Anomalies shown in each panel were calculated relative to the mean over the period 11–0 kyr BP. The dark black line in a shows the mean annual temperature reconstruction adjusted for possible lags in the response of vegetation to climate¹⁷. Horizontal lines and black squares in a represent the observed mean temperature anomaly for the past century, decade and year calculated using the observed NASA GISS Northern Hemisphere land temperatures¹⁰. The thick red line in a represents centennial means of an errors-in-variables reconstruction of Northern Hemisphere temperatures for the past 2,000 years¹ and the orange line shows the observed annual temperatures from GISS since AD 1980¹⁰.

By updating temperature reconstructions from North America and Europe, we evaluate the differences between seasonal and annual trends and the millennial-to-centennial-scale variability in the mid-latitude region that supports the inferred global cooling³ (Fig. 1a, b). Our analyses of fossil pollen data rely on the abundances of more than 40 taxa, each with independent sensitivities to winter cold and growing-season warmth (see Methods). These abundances provide us with the statistical capability to reconstruct mean annual temperatures, the mean temperature of the coldest month of the year and ‘growing degree days’ (a measure of accumulated growing-season warmth, measured in units of kelvin days, K d) using a baseline temperature of 5 °C (ref. 9).

Understanding the trends in these three variables is important not only for understanding Holocene climate dynamics, but also because the warmth of recent decades might have had few precedents in the Holocene if warming indeed dominated northern mid-latitudes for

millennia. To evaluate this possibility, we generated seasonal and annual temperature reconstructions from pollen data using as consistent a process as possible. We used many of the same data as previous syntheses for the individual continents; these previous syntheses used various approaches^{7,8}, but by applying a single approach to an updated data synthesis, we have produced multi-continent reconstructions that document changes that are common across North America and Europe. We compare these reconstructions to a transient CCSM3 simulation. To do so, we used the modern-analogue technique to infer climates from pollen, screened the reconstructions for differences from reconstructions of random variables, gridded the reconstructions to account for spatial biases and used bootstrap resampling to assess uncertainties (see Methods for details). For comparison, we also synthesized and averaged 20 marine and coastal records from our region in the global reconstruction³ (Fig. 1b).

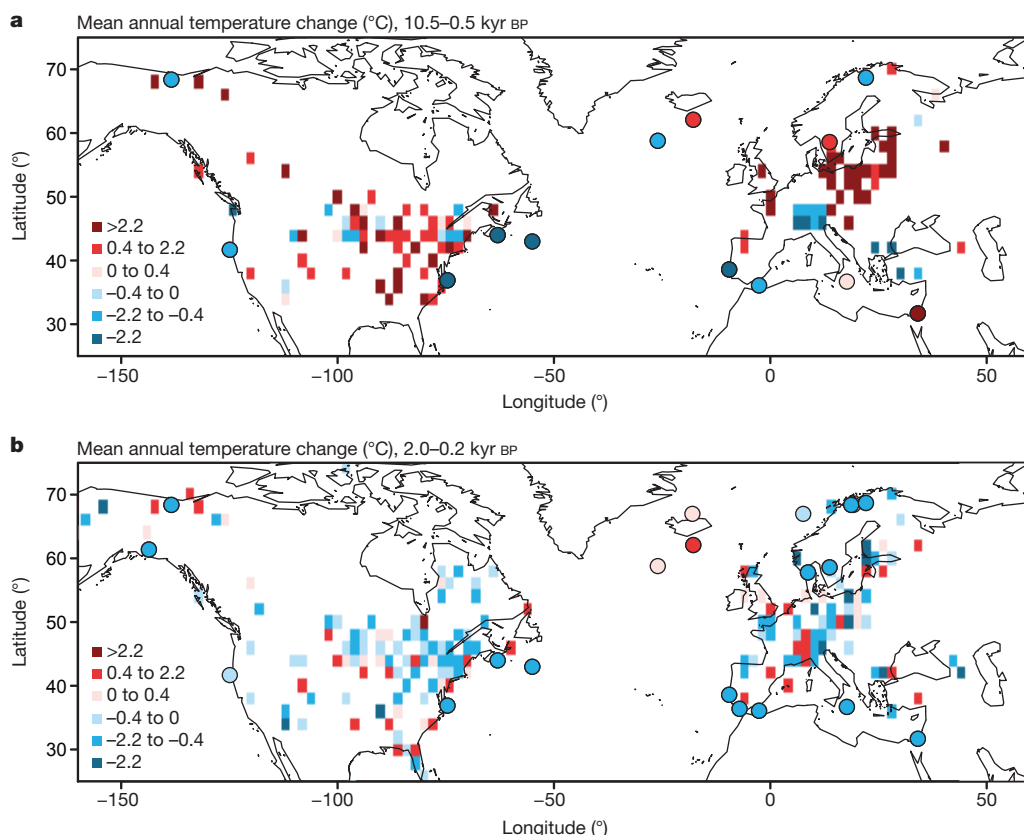


Figure 3 | The geographic extent of the reconstructed temperature trends during the Holocene. **a**, Differences in mean annual temperature between the interpolated values for each grid cell at 10.5 kyr BP and 0.5 kyr BP (value at 0.5 kyr BP less than that at 10.5 kyr BP). Differences denoted by the darkest shades of red and dark blue exceed the root-mean-square error of the modern calibration (2.2 °C). Squares indicate grid cells with data from this study. Circles with black borders indicate records used previously³ and in our marine synthesis (Fig. 1b). **b**, Same as in **a**, but for time slices at 2.0 kyr BP and 0.2 kyr BP (value at 0.2 kyr BP less than that at 2.0 kyr BP).

Comparing the National Aeronautics and Space Administration (NASA) Goddard Institute for Space Studies (GISS) instrumental record¹⁰ with our reconstruction indicates that recent temperatures are above the range of centennial mean temperatures during the Holocene for this region (Fig. 2a). The mean temperature of the past century (AD 1917–2016) was higher than the reconstructed Holocene mean by about 0.3 °C, but was not as high as the warmest centuries of the Holocene (Fig. 2a). However, the mean temperature of the most recent decade (AD 2007–2016), which may be representative of (or probably cooler than) the coming century, exceeds the Holocene mean by approximately 1.2 °C and the mean temperatures of the warmest centuries of the Holocene by more than 0.3 °C even when uncertainties are included (Fig. 2a). We cannot evaluate the range of interannual temperature variations directly, but the mean annual temperature for AD 2016 exceeded all but 70 (0.59%) of the simulated mean annual temperatures for the past 11,000 yr. This finding is important because the CCSM3 simulation reproduces the major trends and magnitudes of changes in our reconstruction and therefore could provide a reasonable estimate of the distribution of annual and seasonal temperatures (Fig. 2).

Consistent with previous estimates from the continents individually^{7,8}, we find that most regions of North America and Europe recorded their lowest, rather than highest³, temperatures early in the Holocene; Holocene warming is a widely replicated trend (Fig. 3a). As indicated by CCSM3⁴ and other simulations¹¹, temperature depression caused by remnant ice sheets extended across North America and Europe. This temperature depression also explains why the early-Holocene temperature anomalies in our reconstruction are lower than in the global mean reconstruction that excludes data from the North Atlantic³ (Fig. 1d, e). Reconstructed growing degree days also rose by more than 200 K d from 11 kyr BP to 5.5 kyr BP (that is, 11–5.5 millennia (kyr) before present (BP), taken to be AD 1950), while winter temperatures rose by approximately 3.75 °C, consistent with the simulated effects of the declining albedo, elevation and meltwater of ice sheets¹¹.

After the disappearance of the ice sheets between 8 kyr BP and 6 kyr BP, the major seasonal trends differ from one another, but

parallel the simulated effects of the dominant climate forcings of the Holocene^{3,4} (Fig. 2a, c). After 5.5 kyr BP, summer cooling decreased the growing degree days by roughly 100 K d as summer insolation declined in the Northern Hemisphere^{4,11,12}, even though the mean temperature of the coldest month continued to increase by 0.3 °C until 2 kyr BP. In contrast to simulated trends forced by greenhouse gases and winter insolation⁴, we reconstruct a winter ‘thermal maximum’ between 5 kyr BP and 2 kyr BP before winters cooled by 0.5 °C to the temperatures of recent centuries (Fig. 2d).

Comparisons with the simulation (Fig. 2) demonstrate that differential variations in seasonal temperature (that is, different trends in different seasons) can explain the differences between the marine and continental reconstructions⁴. Summer temperature sensitivities of plants depend on the integrated effects of maximum warmth and growing-season length, as represented by growing degree days¹³, and pollen-inferred trends in growing degree days closely parallel simulated trends (Fig. 2c). By contrast, many marine and aquatic indicators of palaeo-temperatures have been interpreted to represent maximum summer water temperatures¹⁴, and the marine synthesis similarly tracks the simulated June–August mean temperatures over land, which reached a maximum early in the Holocene (Fig. 2b).

Growing degree days increased over the Holocene even as June–August temperatures declined because, as summer insolation anomalies decreased, the effects of cool-season insolation and atmospheric greenhouse gas concentrations increased⁴ sufficiently to lengthen the growing season and offset the decline in maximum warmth. Therefore, although the marine- and pollen-based reconstructions (Fig. 2b, c) may both relate to the warm season (as variously defined), the simulation shows that different aspects of the warm season probably had different trends. Differences among warm-season sensitivities may also explain other paradoxes among aquatic and terrestrial records¹⁵. However, the explanation here is imperfect: summer cooling simulated over the land (Fig. 2b) does not extend to summer temperatures over the ocean as simulated by CCSM3 (Extended Data Fig. 1).

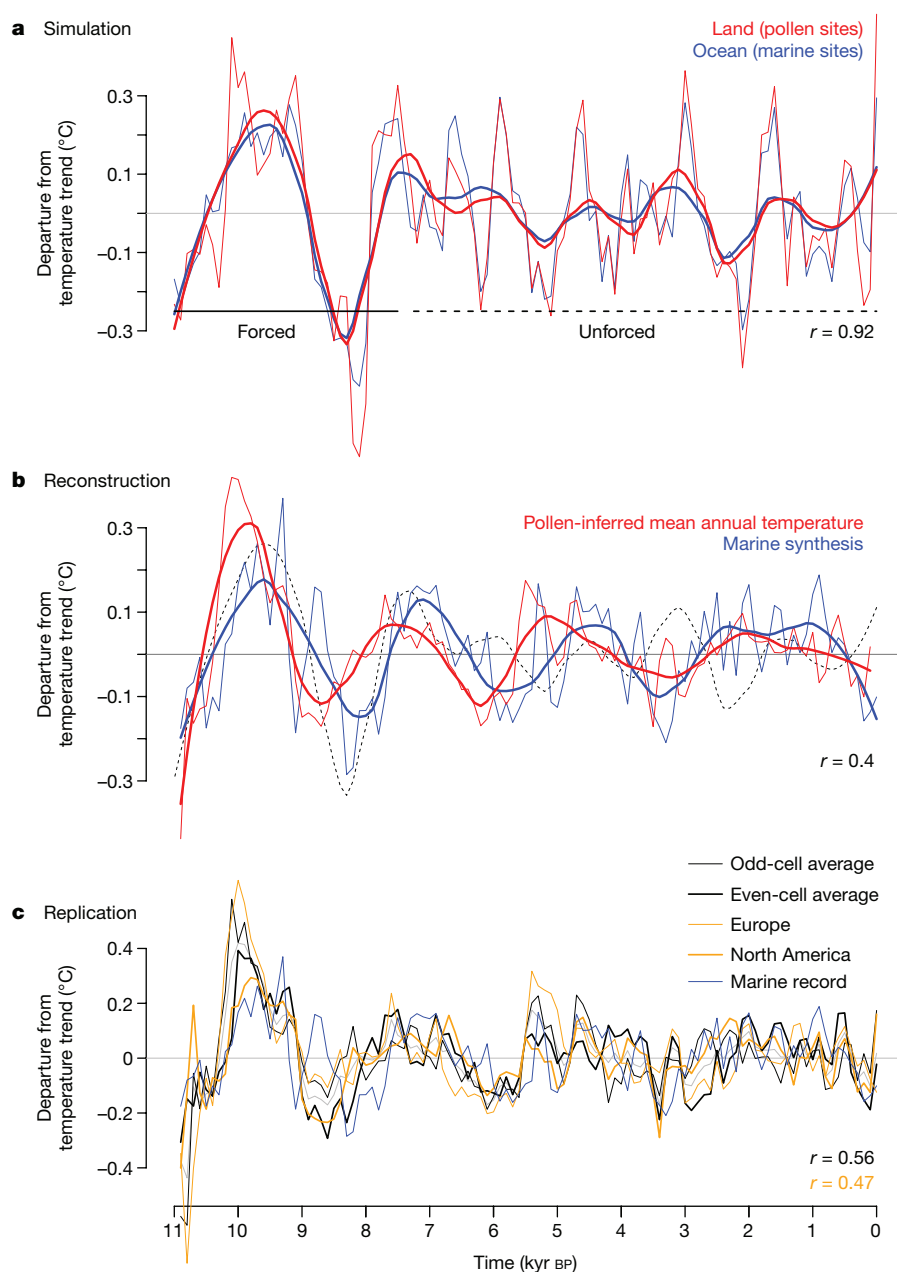


Figure 4 | Simulated and reconstructed departures from long-term temperature trends, including forced and unforced variations on centennial scales. **a**, The simulation demonstrates that centennial-mean temperature variations (thin lines) over land (red) and the ocean (blue) could have been well correlated. Thick lines show 2,500-yr locally weighted regression fits. **b**, Departures from the long-term trends in the continental (red) and marine (blue) temperature reconstructions are consistent with the simulated patterns in **a** (the long-term trends are shown in Extended Data Fig. 1); thick and thin lines show 100-yr means and 2,500-yr fits, as in **a**. The thick red line from **a** is shown as the dashed black line. **c**, The de-trended mean annual temperature reconstructions also correlate across continents (orange: North America, thick; Europe, thin) and when the reconstructions are averaged across all even- or odd-numbered cells on both continents (black: even, thick; odd, thin). The de-trended marine synthesis reconstruction is shown in blue. For comparison, the thin grey line shows the mean pollen-inferred reconstruction for all locations (red thin line in **b**). Inset numbers indicate the correlation coefficients r . All departures are relative to the long-term trends determined using 6,500-yr locally weighted regression fits to the simulation and reconstructions (Extended Data Fig. 1).

Regardless of the specific explanation, however, data-model divergence is not apparent between the pollen data and the simulation of the northern mid-latitudes, and is therefore confined to North Atlantic records (Fig. 1a, b). The terrestrial reconstructions match simulated warming trends that extend to 5.5–2.1 kyr BP depending on the season. Cooling became well expressed across seasons only in the past two millennia or so (Fig. 2a). The marine synthesis (Fig. 1b) thus differs from the simulation of mean annual temperature (Fig. 2a), our terrestrial reconstructions (Fig. 1e) and the global average outside the North Atlantic region (Fig. 1d) because alkenone-inferred temperatures from the western Atlantic indicate cooling of more than 5 °C since the Younger Dryas interval^{3,7} (Fig. 1a). This cooling contrasts with nearby coastal pollen records, as well as chironomid and isotopic records, which indicate a cold Younger Dryas interval¹⁶ and subsequent warming (compare trends in Fig. 3a). Consistent with summer insolation effects on maximum temperatures even during the Younger Dryas interval, the mean difference between the marine and continental reconstructions has progressively decreased from a maximum of more than 3 °C at the start of the Holocene^{3,4}, but nutrient or other non-climatic changes could have also influenced the marine records.

The simulation further supports the hypothesis that multi-century variation was important during the Holocene. Coherent changes with durations of 100–1,000 yr affected large-area averages of ocean and land temperatures in the simulation (Pearson's product moment correlation coefficient $r = 0.92$, Fig. 4a). Similarly, the continental and marine reconstructions also correlate on centennial-to-millennial scales ($r = 0.40$, Fig. 4b). Although local-to-regional factors prevent the mean patterns from appearing in all individual records (see, for example, Fig. 1a), correlation on millennial scales appears for both continents (Fig. 4c, orange lines) and is retained when we divide the data into groups (Fig. 4c, black lines). The correlation also exceeds the range of potential correlations produced using 1,000 random autoregressive series of the same order as the reconstructions (95% range: $r = -0.25$ to $r = 0.25$; Extended Data Fig. 1). Continental temperature departures from the long-term trends typically lead the temperature anomalies in the oceans (Fig. 4b), but multiple factors, including seasonal sensitivities, dating uncertainties and ocean–land thermal differences, could explain the relationship.

Comparison of the simulation and reconstructions indicates four distinct modes of variation that are similar in magnitude on centennial scales in the reconstructions (about 0.75 °C) and in the simulation

(about 0.6 °C). The first mode arises from early-Holocene ice-sheet and meltwater dynamics, which produce changes such as low temperatures across seasons from 9 kyr BP to 8 kyr BP (Fig. 2). Similarities between the reconstructions (including previous work⁷) and the simulated climate reflect strong forcing in the early Holocene (Fig. 4a, b). Some differences relate to how the forcing was imposed in CCSM3⁴; but, as in the reconstructions, simulated changes in ice-sheet configuration and meltwater routing reversed the long-term warming trend at around 9–8 kyr BP. Related changes accelerated warming at 10 kyr BP and 8.2 kyr BP (Fig. 4).

The second mode arises from unforced, multi-century temperature variations in the simulation, which have similar durations and magnitudes (± 0.2 °C) to those in the reconstructions (Fig. 4a). The simulation did not include solar irradiance or volcanic forcing, only long-term controls such as orbitally forced insolation anomalies, greenhouse gases and ice-sheet topography⁵. Therefore, the simulated variations on centennial scales after around 8 kyr BP cannot be ascribed to external forcing in the model (Fig. 4a). Differences in the timing of simulated and reconstructed variations since 8 kyr BP (Fig. 4b) are consistent with stochastic ocean–atmosphere dynamics, which would manifest as similar changes, but at different times in different realizations, such as between multiple model runs or between the model and the ‘real’ climate revealed by the reconstructions (for example, cooling at 4–3 kyr BP in the reconstructions (Fig. 4b) and at 3–2 kyr BP in CCSM3 (Fig. 4a)).

Different reconstructions from a given region, however, should have recorded the same events at the same time even if they arose stochastically. We find agreement between the marine- and pollen-based records (Fig. 1d, e) as well as with other records for the past two millennia (Fig. 2a). For example, our reconstruction for the Little Ice Age correlates with the 100-yr mean of an error-in-variables reconstruction for the Northern Hemisphere¹ (Fig. 2a, red line). If we account for lags in vegetation responses to climate based on forest succession timescales of around 300 yr (ref. 17), then we find greater similarity between the magnitudes of simulated and reconstructed changes (Fig. 2a, thick black line).

The third mode is a rapid, mid-Holocene temperature increase at 5.5 kyr BP, which was largest in the reconstructed growing-season temperatures (growing degree days, Fig. 2c). This rapid increase may relate to abrupt shifts in remote regions, such as the rapid onset of aridity and coastal upwelling in western Africa, which took place at the same time¹⁸. These changes could represent thresholds or internal feedbacks, which are thought to be important in the tropics, but which probably had consequences that extended to the Arctic^{18,19}. Other models explicitly produce such a state shift^{18,19} and underscore the potential for unexpected state changes even when external forcing is slowly varying. The fourth mode is the winter cooling after 4 kyr BP (ref. 20), which accelerated over the past two millennia (Fig. 2d). This mode could represent the importance of negative feedbacks, interactions between seasonal trends (such as reduced winter heat transport to the extra-tropics via sea-ice expansion in response to summer cooling) and unconstrained external forcing (such as volcanism)¹.

Overall, our reconstructions indicate that the on-going warming today would have started from a baseline approximately 0.5 °C higher than observed had millennial-to-centennial-scale variations not produced cooling over the past two millennia that deviated from Holocene trends²¹. The reconstructions support the ability of models such as CCSM3 to capture large-scale climate responses to external forcing and important internal dynamics. Although additional transient climate simulations and new, detailed palaeoclimate records are needed to further understand the processes involved, millennial-to-centennial-scale climate variability such as occurred during the Holocene could continue to amplify or modulate future temperature trends.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 February; accepted 5 December 2017.

- Mann, M. E. *et al.* Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science* **326**, 1256–1260 (2009).
- PAGES 2k Consortium. Continental-scale temperature variability during the past two millennia. *Nat. Geosci.* **6**, 339–346 (2013).
- Marcott, S. A., Shakun, J. D., Clark, P. U. & Mix, A. C. A reconstruction of regional and global temperature for the past 11,300 years. *Science* **339**, 1198–1201 (2013).
- Liu, Z. *et al.* The Holocene temperature conundrum. *Proc. Natl Acad. Sci. USA* **111**, E3501–E3505 (2014).
- Liu, Z. *et al.* Transient simulation of last deglaciation with a new mechanism for Bølling–Allerød warming. *Science* **325**, 310–314 (2009).
- Sachs, J. P. Cooling of Northwest Atlantic slope waters during the Holocene. *Geophys. Res. Lett.* **34**, L03609 (2007).
- Viaou, A. E., Gajewski, K., Sawada, M. C. & Fines, P. Millennial-scale temperature variations in North America during the Holocene. *J. Geophys. Res.* **111**, D09102 (2006).
- Mauri, A., Davis, B. A. S., Collins, P. M. & Kaplan, J. O. The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation. *Quat. Sci. Rev.* **112**, 109–127 (2015).
- Bartlein, P. J. *et al.* Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis. *Clim. Dyn.* **37**, 775–802 (2011).
- Hansen, J., Ruedy, R., Sato, M. & Lo, K. Global surface temperature change. *Rev. Geophys.* **48**, RG4004 (2010).
- Renssen, H. *et al.* The spatial and temporal complexity of the Holocene thermal maximum. *Nat. Geosci.* **2**, 411–414 (2009).
- Berger, A. & Loutre, M. F. Insolation values for the climate of the last 10 million years. *Quat. Sci. Rev.* **10**, 297–317 (1991).
- Prentice, I. C. *et al.* A global biome model based on plant physiology and dominance, soil properties and climate. *J. Biogeogr.* **19**, 117–134 (1992).
- Lorenz, S. J., Kim, J.-H., Rambu, N., Schneider, R. R. & Lohmann, G. Orbitally driven insolation forcing on Holocene climate trends: evidence from alkenone data and climate modeling. *Paleoceanography* **21**, PA1002 (2006).
- Samartin, S. *et al.* Warm Mediterranean mid-Holocene summers inferred from fossil midge assemblages. *Nat. Geosci.* **10**, 207–212 (2017).
- Shuman, B. N. & Marsicek, J. The structure of Holocene climate change in mid-latitude North America. *Quat. Sci. Rev.* **141**, 38–51 (2016).
- Webb, T. Is vegetation in equilibrium with climate? How to interpret late-Quaternary pollen data. *Plant Ecol.* **67**, 75–91 (1986).
- deMenocal, P. *et al.* Abrupt onset and termination of the African Humid Period: rapid climate responses to gradual insolation forcing. *Quat. Sci. Rev.* **19**, 347–361 (2000).
- Muschitiello, F., Zhang, Q., Sundqvist, H. S., Davies, F. J. & Renssen, H. Arctic climate response to the termination of the African Humid Period. *Quat. Sci. Rev.* **125**, 91–97 (2015).
- Walker, M. J. C. *et al.* Formal subdivision of the Holocene Series/Epoch: a discussion paper by a Working Group of INTIMATE (integration of ice-core, marine and terrestrial records) and the Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy). *J. Quat. Sci.* **27**, 649–659 (2012).
- Schurer, A. P., Mann, M. E., Hawkins, E., Tett, S. F. B. & Hegerl, G. C. Importance of the pre-industrial baseline for likelihood of exceeding Paris goals. *Nat. Clim. Change* **7**, 563–567 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Webb, S. Marcott, J. Shinker, T. Minckley, B. McElroy and E. Currano for comments. Data were obtained from the Neotoma and European Pollen Databases, and we acknowledge the work of the data contributors and the Neotoma and European Pollen Database communities. TraCE-21ka was made possible by the DOE INCITE computing programme, and supported by NCAR, the NSF P2C2 programme, and the DOE Abrupt Change and EaSM programmes. Funding was provided by Wyoming NASA Space Grant (#NNX10AO95H) and EPA STAR (FP-91763201-0) to J.M., and NSF support to B.N.S. (DEB-1146297), S.B. (EAR-1003848) and P.J.B. (ATM-06202409). S.L.S. was supported by the US Geological Survey Climate Research and Development Program.

Author Contributions J.M. and B.N.S. oversaw and contributed to all aspects of the research and, with S.B., designed the project. P.J.B. and S.L.S. contributed analyses of the CCSM3 climate model simulation. J.M. and B.N.S. carried out the analyses and wrote the first version of the paper, and J.M., B.N.S., P.J.B. and S.L.S. contributed to the final version.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.M. (jmarsicek@wisc.edu).

Reviewer Information Nature thanks M. Haran and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Fossil pollen datasets. Our analysis is based on climatic interpretation of fossil pollen stratigraphies from lakes and wetlands across North America and Europe. The fossil pollen datasets included meet the following criteria: (1) they contain to samples within the time window from 11.0 kyr BP to present; (2) they are publicly available from the European Pollen Database²² (EPD; <http://www.europe.anpollendatabase.net>) or the Neotoma Paleocology Database²³ (<http://www.neotomadb.org>); and (3) they contain original sampling depths, age models and pollen counts for handling chronologies and carrying out the modern-analogue technique (MAT).

The fossil pollen datasets were extracted from the databases in July 2013. Applying our criteria yielded 1,605 fossil pollen records from North America ($n = 799$) and Europe ($n = 806$). Age controls for the records were derived primarily from radiocarbon dating of lake sediments, core tops, biostratigraphic dates (pollen horizons), varved lake sediments and ²¹⁰Pb dates. Extended Data Fig. 2 provides a flow chart of the reconstruction process with information about the data files and R code used.

Chronologies. We used the best-available calibrated radiocarbon chronologies for each site, and focus on chronologically robust aspects of the reconstructions. Chronologies from the databases were carried over with the pollen count data for each site. Chronologies from the EPD were recently converted to calendar years²², but chronologies from Neotoma were obtained in radiocarbon years. Neotoma (that is, North American) chronologies were therefore updated and calibrated using revised age models^{24–26}. For the sites that did not have updated chronologies, we used the calendar year calibration curve in Neotoma²³ to generate chronologies. **MAT.** Our approach relies on the assumption that pollen taxa (and the plants they represent) are influenced not only by growing-season temperatures, but also by minimum winter temperatures that limit plant growth, regeneration and thus distributions¹⁴. Pollen-inferred reconstructions generated under this assumption consistently agree with independent indicators^{27,28}, accurately reproduce modern spatial patterns of key climate variables^{29,30}, and have the statistical power (more than 40 taxa as predictor variables) and spatial density to constrain spatial and seasonal uncertainties^{9,29}. We exploit these strengths to reconstruct temperature changes during multiple seasons, including in mean annual temperature (AnnT), in growing degree days using a 5°C baseline (GDD5) and in the mean temperature of the coldest month (MTCO).

To do so, we applied the MAT^{29,31} to generate temperature reconstructions from all of the available fossil pollen records. The technique matches fossil and modern pollen samples using the squared chord distance (SCD) dissimilarity metric³¹ to infer past temperatures from those associated with the modern samples that have taxonomic assemblages most like each fossil assemblage. The comparisons rely on similar, but continent-specific, lists of taxa (Supplementary Table 1) and extensive modern pollen datasets^{30,32}.

Modern pollen–climate calibration. We used a modern pollen dataset derived from previous work as the basis for analogue matches to the fossil data. The modern dataset consists of 4,833 surface samples from North America³⁰ and 3,731 samples from Europe³². Modern climate data^{33,34} have previously been assigned to each modern surface sample via interpolation^{30,32} and, following ref. 9, we use these climate values as the basis for our inferred climates.

Taxonomic treatment of the pollen data. We used similar multivariate pollen datasets for each continent, but base our statistical reconstructions on 60 and 40 taxa for North America²⁹ and Europe³⁵, respectively, as recommended in previous studies^{29,35}. Because many tree genera contain distinct sets of species between eastern and western North America and are differently distributed along climate gradients, we split morphologically indistinguishable pollen types from the relevant taxa into eastern and western groups (and northeastern and southeastern groups for *Pinus*), consistent with recommendations²⁹. Regional splits were applied to *Abies*, *Acer*, *Alnus*, *Celtis*, *Cupressaceae*, *Fraxinus*, *Larix*, *Picea*, *Pinus*, *Quercus* and *Tsuga*. The regional groups of species within these morphological types of pollen currently overlap only minimally, if at all, and were unlikely to have overlapped over the past 11,000 years. No regional splits were recommended³⁵ for fossil and modern pollen data in the EPD taxa. We used the existing taxonomic hierarchy within Neotoma to combine pollen types for some higher-level groups, such as *Ericaceae*, when extracting the fossil pollen data for North America²³, but we recognize that this scheme may differ from the choices of some individual investigators at the site level.

We used the R package *palaeoSigs*³⁶ to sum the raw pollen counts for each sample in the fossil and modern pollen datasets and to calculate percentages for each individual taxon based on the sum of the total counts of all of the taxa in our taxa lists (Supplementary Table 1). We then used the percentages to carry out modern validation of the MAT before applying the technique to the fossil data³⁶.

Modern validation experiments. To test how well the MAT predicts observed modern temperatures using the modern surface sample dataset, we conducted leave-one-out analyses²⁹. The analysis predicts the modern climate associated with

each modern surface sample using the best analogues from the modern pollen database consistent with the recommendation for fossil samples²⁹. We use the leave-one-out analyses to evaluate the number of analogues to include as the basis for reconstruction; see below. By definition, the analogues exclude the modern sample that is being reconstructed. Each modern pollen assemblage was compared to all of the remaining modern pollen data using the SCD dissimilarity metric, which we applied using the MAT function in the R package *rioja*. The MAT was applied through the *palaeoSigs* package for fossil analyses^{36,37}. The MAT applies the following equation³⁸ to compare each fossil and modern sample:

$$d_{ab} = \sum_c \left(p_{ac}^{0.5} - p_{bc}^{0.5} \right)^2 \quad (1)$$

where d_{ab} is the SCD dissimilarity between pollen samples a and b , p_{ac} is the percentage of pollen of type c in sample a , the target sample, and p_{bc} is the percentage of pollen of type c in sample b , a possible analogue.

A good analogue has a low SCD; zero SCD indicates that two samples are identical in their pollen percentage composition. When we used more than one analogue, the best analogues were averaged to produce a climate estimate for the location of the modern pollen assemblage. To measure the accuracy of the reconstruction, we used linear regression to measure the fit between the modern observed and reconstructed climate values across the entire modern dataset and generate an R^2 and root-mean-square error (RMSE) for each variable (Extended Data Table 1).

Experiments on number of analogues and data gridding. We tested whether the modern calibration RMSE was improved by gridding the data and averaging over multiple analogues (for example, the seven best²⁹). Both approaches involve averaging multiple samples, either across fossil sites (that is, averaging the climates associated with the analogues of multiple fossil samples) within a grid cell or across analogues for each individual fossil sample. We use the term ‘gridding’ to refer to calculating the mean of the temperatures reconstructed from one or more pollen sample within evenly spaced spatial windows every 100 years.

Gridding can help to address spatial biases in the fossil dataset⁵, while also amplifying the climate signal relative to other sources of palaeo-ecological variability. Local ecological factors, such as soils, fires or competition, may alter the climate–vegetation relationship at a particular site, making a given site a poor indicator of the local-to-regional climate over some window of time (for example, after a fire). Some sites may have a persistently weak link to climate (see soil effects at sites such as Duck Pond, Massachusetts, USA²⁸), but other sites may be only temporarily affected by disturbances and other factors. Because sites in the same region will have different local disturbance and edaphic histories, but the same climate history, averaging reconstructions from many sites per grid cell can reduce the local ecological noise and provide a more robust estimate of the local-to-regional climate signal. Averaging multiple analogues per fossil sample may similarly improve the accuracy of the reconstruction because a large number of analogues reduces stochasticity and improves precision in MAT reconstructions.

Alternatively, choosing the single best analogue can avoid blurring climate signals through averaging and can reduce errors produced by combining information from a wide range of potentially weak analogues⁷. However, the best analogue may result from chance (for example, interactions among uncertain pollen counts) rather than from climatic determinants²⁹. Likewise, because the modern pollen data represent a finite set of climate conditions, use of a single analogue can artificially quantize reconstructions with flickering or steps between samples even when the climate history was stationary or smoothly varying.

For these reasons, we tested whether gridding, using either the best analogue for each sample⁷ or the mean of the seven best analogues²⁹, improved the accuracy of the reconstruction. To do so, we carried out experiments for AnnT, GDD5 and MTCO using the modern pollen–climate dataset to contrast the effects of gridding and averaging analogues on the RMSE (Extended Data Table 1).

To produce the gridded datasets, we matched coordinates from modern site locations to their corresponding cell within a $2^\circ \times 2^\circ$ grid, which we also apply (on the basis of these results) to the fossil data (using the R code *GriddingCode_Marsicek.R*). The $2^\circ \times 2^\circ$ grid is similar to that used in other hemispheric-scale analyses^{1,9} and typical of palaeoclimate model output. Using a 2° grid-cell size avoids over-smoothing the regional climate patterns⁹. The grid extends from 15° N to 80° N in 2° increments and from 170° W to 60° E in 2° increments, but we calculate climate values only for cells with one or more pollen sample. The grid-cell value (such as the modern value used in these experiments) was calculated by averaging the analogue climates for all pollen samples in the grid cell. Because each surface sample within a grid cell had both an observed modern temperature (AnnT, GDD5, MTCO) and a reconstructed temperature based on the selection of the best analogue or the seven best modern analogues, we gridded and compared the observed and predicted values using the same approach.

To account for statistical effects produced by differences in the total number of pollen samples versus the total number of grid cells, we also conducted experiments that reduced the number of sites to the number of grid cells by randomly choosing a modern sample from each grid cell (Extended Data Table 1). By doing so, we could compare the effects of gridding to the effects of using more than one analogue per sample directly without biases related to different numbers of samples.

Overall, using the seven best analogues for AnnT consistently improved the accuracy of the reconstruction skill by more than 1.9% compared to using only the best analogue (for example, by 9.6% without also gridding, a RMSE of 2.09 °C versus 2.29 °C). Gridding reduces the RMSE from the 2.78–2.91 °C associated with using only one representative sample per grid cell to 2.19–2.23 °C when all sites are averaged per grid cell. (Direct comparison of all non-gridded to gridded results is confounded by the difference in the number of samples involved: 8,564 individual samples versus 820 grid cells.) R^2 also increases from 0.90–0.91 to 0.94–0.95 (Extended Data Table 1).

On the basis of these experiments, we decided to grid the reconstructions and to use the seven best analogues. Using these criteria provided RMSE values of 2.2 °C for AnnT, 410 K d for GDD5 and 3.3 °C for MTCO. Because many grid cells contain only a single fossil record, these methodological choices reduced the possibility of false positives and artificially quantized reconstructions that result from using only one analogue^{29,38}. Because a wide range of human land use (for example, intensively affected areas of Europe and eastern North America and boreal areas with limited effects) affected the modern pollen data, the calibration results also provide support for applying the method to past periods with a range of different human effects, especially after gridding the data to limit the influence of local ecological signals.

MAT and significance testing of the reconstructions. Once we decided which methods to apply to maximize reconstruction accuracy, we applied the MAT to the fossil data using palaeoSig³⁶. By doing so, we also followed a recommendation³⁹ to examine whether a given climate variable can be reconstructed from fossil pollen at a specific site (a relationship existing between climate and pollen assemblages today does not necessarily mean that the relationship has any predictive value for the past).

To ensure that records contained robust signals of both long-term trends and additional (but possibly weaker) shorter-term temperature variability, we examined whether the reconstruction from each site was significantly different from 99 reconstructions of random, autocorrelated environmental variables (that is, the signals carried more information than the random variables). Following ref. 39, we first used palaeoSig to estimate the proportion of variance explained by the fossil pollen data at each site using redundancy analysis, a type of constrained ordination. Then, we generated reconstructions of synthetic randomly generated variables using the MAT. To do so, spatially autocorrelated fields of the random variables were drawn from a uniform distribution, and the values from these fields were assigned to the modern pollen dataset in place of the observed modern climate variables. equation (1) was applied and the seven best analogues were averaged, using the same process as would be applied to reconstructions of the observed climate variables, to infer time series of the random variables. Next, we used the 99 reconstructions of random variables to produce a null distribution of variance explained in each fossil pollen record using the redundancy analysis results. The reconstructions of the climate variables using the MAT (AnnT, GDD5 and MTCO) were then flagged as 'significant' if they explained more of the variance in the fossil pollen data than did 95% of the random reconstructions. The significance values represent the percentage of the random variables that explain more of the variance in the fossil pollen data than do the actual temperature reconstruction^{36,39}.

Fewer than 300 reconstructions of AnnT, GDD5 or MTCO explained more of the variance in the associated fossil pollen record than did 95% of the random reconstructions. However, because the test is conservative (for example, it inherently fails for sites with no trends, such as at climatic hinge points, and assumes a linear relationship to the components of the pollen record extracted through redundancy analysis) and substantially reduced the spatial coverage of reconstructions, we compromised between accepting reconstructions that were falsely considered significant (type I errors) and increasing the spatial array of reconstructions used here. To do so, we increased the number of records used by including all temperature reconstructions that explained more variance in the pollen data than did the mean plus one standard deviation (84.2%) of the 99 reconstructions of random dummy variables. Of the 1,605 sites, this step retained 565 reconstructions for the AnnT reconstruction (35.2%, Extended Data Fig. 3a), 584 for GDD5 (36.4%, Extended Data Fig. 3b) and 579 for MTCO (36.1%, Extended Data Fig. 3c). Overall, at least one variable at 843 sites was considered significant by this metric, but the major patterns in the composite reconstructions differ little even if all site-specific reconstructions are included in the mean (as shown in Extended Data Fig. 5). 642 of the sites with significant reconstructions contained data younger than 0.5 kyr BP, as required for us to calculate our mean time series on the basis of reconstructed

temperature anomalies relative to a consistent baseline using mean temperatures since AD 1450. For the time series of AnnT, GDD5 and MTCO (Extended Data Fig. 3a–c), 415, 453 and 427 sites, respectively, contained data younger than 0.5 kyr BP, as required to calculate the anomalies (Supplementary Table 2).

Of the sites with a significant variable, only 52 (about 6%) did not have updated chronologies, so we used the calendar-year calibration curve in Neotoma²³ to generate chronologies. Also, the random removal of significant sites with high SCD values (more than 0.3 average for all values) did not substantially affect the reconstruction. The total numbers of samples and age control points (such as ¹⁴C ages) used per century for this final set of reconstructions are provided in Extended Data Fig. 4. The region most affected by the screening for significance is the boreal forests of the two continents (that is, western Canada and northern Scandinavia; Extended Data Fig. 3).

Temporal interpolation of the data. Gridding the fossil reconstructions required the reconstructions for each site have consistent time steps. Therefore, once we generated and identified the significant reconstructions for AnnT, GDD5 and MTCO, we interpolated each one to pseudo-centennial (100-yr) time steps. The interpolation was based on the median age model for each site and therefore may only approximate the true age of the reconstructed sample. We discuss age uncertainty effects below, but use 100-yr interpolated time steps for averaging and mapping, which is consistent with other studies^{7,28}.

As noted above, we also converted the temperature reconstructions to anomalies from the mean of the 500 years before present (AD 1450–1950) in the interpolated record. We applied this approach in part because not all records contained a modern sample. The mean values of the 500 years before present also capture the pre- and post-industrial time period (around AD 1850) and represent a compromise from studies that use 'core tops' and pre-industrial baselines to calculate the temperature anomalies. Because many cores were collected before AD 2000 and because pollen data act as a low-pass filter on climate signals via processes such as tree longevity and sediment mixing, we did not expect to reconstruct, nor did we detect, the increase in temperatures over recent decades. For comparison with the simulation, which generated higher temperatures since 0.5 kyr BP than our mean reconstructions but otherwise correlated closely with our reconstructions (see, for example, Fig. 2a), we also plot the mean time series relative to their Holocene means in Fig. 2 by calculating the mean departure from AD 1450–1950 for the whole Holocene and then subtracting this mean departure from each time series.

Comparison data. We compared the AnnT time series generated from the gridded reconstructions (see below) with the GISS composite of observed Northern Hemisphere land temperatures¹⁰. We subtracted the mean for AD 1900–1999 from the observed temperatures and added the resulting anomalies to the 0 yr BP (AD 1950) value in our AnnT reconstruction to ensure direct comparability to our Holocene baseline. We further compare the reconstructions with both the centennial means of an error-in-variables reconstruction of Northern Hemisphere temperatures for the past 2,000 years¹ (red line, Fig. 2a) and the mean of the northern mid-latitude records used in the existing synthesis of global mean temperatures³ (Fig. 1b). To do the latter, we extracted those records ($n = 20$; Extended Data Table 2) contained within our spatial domain, extending from before 11 kyr BP to after 0.5 kyr BP, and averaged them using the same approach that we applied to our data, including calculating anomalies relative to the means for AD 1450–1950. We refer to this new average as the marine synthesis because it derives primarily from marine geochemical records.

Mean time series and uncertainty analysis. Once the data were interpolated and gridded, we calculated mean time series for all three temperature variables on the basis of all grid cells containing significant reconstructions across the two continents. The long-term trends in the means of the gridded reconstructions (such as the dark red line in Extended Data Fig. 5) do not differ substantially from the means of all reconstructions, even when including records that were not identified as significant (such as the blue line in Extended Data Fig. 5). However, spatial biases induced by clusters of records (such as the high density of records in mid-latitude eastern North America) cause the gridded means to be lower than the non-gridded mean (Extended Data Fig. 5). The gridding amplifies the millennial-scale variability because the millennial signals are not damped by oversampling of regions and grid cells not representative of the overall trend (as in the blue line in Extended Data Fig. 5). In the main text, we emphasize the multi-century features of the reconstructions that appear both with and without screening the data before gridding (such as temperature maxima at around 9 kyr BP and 5.5 kyr BP).

We used a bootstrap resampling approach to assess the uncertainty in our mean reconstructions. To do so, we randomly removed half of the grid cells 100 times to produce 100 different versions of the time series for each temperature variable. The bootstrapping allowed us to evaluate the biases produced because of both the age uncertainties (by chance, some sites have median sample ages assigned by the best-available age models that were either erroneously young or old) and the

temporal density of individual pollen samples (because of the potential for aliasing the underlying signals). Our time series for each reconstruction derives from the median of the 100 bootstrapped iterations of the mean of all grid cells (dark red line, Extended Data Fig. 6).

To further quantify the effects of uncertainty related to the MAT, we created a second ensemble of 100 bootstrapped iterations by again randomly removing 50% of the grid cells (thin red lines, Extended Data Fig. 6), but then adding noise to each individual grid cell by drawing from a normal distribution with a standard deviation equal to the RMSE of our modern calibrations before averaging to produce our summary time series^{3,40} (Extended Data Table 1; using the R code `UncertaintyAnalysisCode_Marsicek.R`). The second ensemble (red band denotes the 2.5%–97.5% quantiles of the 100 individual iterations with noise, Fig. 1, Extended Data Fig. 6) enabled us to assess the effect of using reconstructions degraded by reconstruction error; because the random noise added to each reconstruction is not spatially autocorrelated (like it is for climate and vegetation), it estimates the range of uncertainty conservatively. Both ensembles produce similar uncertainty distributions (Extended Data Fig. 6) even though the second ensemble includes all of the steps of the first plus the additional signal degradation; the limited change results from cancelling of the noise across many cells during the averaging process.

Diagnosing the millennial-scale variability. To evaluate the robustness of the millennial-scale signals, we evaluated correlations with the marine synthesis and among subsets of the data. To do so, we fitted a 6,500-yr locally weighted regression ('loess')⁴¹ and subtracted it from the pollen-based AnnT reconstructions to de-trend the data (Extended Data Fig. 1). However, de-trending approaches can influence the residuals and loess does not include an explicit autoregressive term. Therefore, we also applied a generalized additive mixed model (GAMM) fit to each series as a function of time, including an autoregressive term (Extended Data Fig. 7a; using the R code `TimeSeries_Decomposition_Marsicek.R`). The loess and GAMM fits do not differ meaningfully from each other. Once the long-term trends are removed from the data, we find correlations in the millennial variability between the continental and marine syntheses (Fig. 4b), and between North America and Europe ($r=0.75$; Extended Data Fig. 7d). Because we use the same de-trending approaches for both datasets, the residual patterns should be comparable even if they would shift modestly using different approaches. The correlations are greater than 95% of those between randomly generated autocorrelated time series (Extended Data Fig. 1).

Wavelet analysis is also ideal for detecting changes in time series data because it can deal with quasi-periodic signals and varying or noisy ones⁴². We performed wavelet analysis in R on the de-trended marine and continental records using the `xwt` function in the `biwavelet` package⁴³. We set the spacing between the scales to 1/24, the length of the time steps to 100 and the significance level to 0.90, and used a Morlet waveform as the mother wavelet (see Extended Data Fig. 8). We also calculated cross-wavelet significance using the `xwt` function in the `biwavelet` package, which shows that the early Holocene variations were particularly well correlated across datasets and regions.

TraCE simulation and calculation of AnnT, MTCO and GDD5. To evaluate our reconstruction against the simulated effects of the known climate forcing, we use calendar-corrected output from a transient simulation of annual and seasonal temperatures in the northern mid-latitudes over the past 11,000 years produced by a coupled ocean–atmosphere model, CCSM3⁶. In Fig. 2, we compare the pollen-derived reconstructions of AnnT, GDD5 and MTCO generated here with those obtained from the TraCE-21ka transient climate-model simulations^{4,5,44}. The TraCE-21ka simulations were performed using CCSM3, a fully coupled ocean–atmosphere general circulation model, which was forced by time-varying insolation, greenhouse gases, ice-sheet topography, land/ocean palaeogeography and meltwater fluxes to the ocean, over the interval 22 kyr BP to present (for further details, see <http://www.cgd.ucar.edu/ccr/TraCE/>). We used the output from the full TraCE simulation, available at <https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm3.trace.html>.

Several steps of data reduction and analysis were implemented to calculate AnnT, GDD5 and MTCO from the model-output variable TREFHT (near-surface air temperature). The TraCE model output is available in time steps of months, aggregated using a perpetual present-day 'noleaps' calendar, in which all years are 365 days long and month lengths are defined using the present-day calendar (that is, January is always 31 days long, February is always 28 days long, and so on; see <http://cfconventions.org>). It is well known, however, that Earth's orbital variations cause changes in month lengths that follow the precessional orbital element that governs the times of year of perihelion and aphelion⁴⁵. In particular, months (defined, for example, as the interval over which one-twelfth of an orbit is completed) are shorter when the Earth is at perihelion and longer when at aphelion. In practice, around 10 kyr BP, when perihelion occurred in the boreal summer

(114 days after the vernal equinox, as compared to 76 days before at present), the part of the orbit conventionally labelled 'July' was about two days shorter than present and January was about two days longer⁴⁶. More importantly, these 'calendar effects' also influence when during the year a particular month or season begins or ends⁴⁷. For example, at 10 kyr BP, July began 98 days after the vernal equinox, in contrast to 101 days at present. The effects of summarizing the climate-model output using the present-day definition of months as opposed to one appropriate for a particular time period can be as large as the long-term mean differences in simulated climate between past and present^{45,48}.

There are several approaches for adjusting monthly and seasonal values from the present-day calendar to a calendar appropriate for a particular time^{48–50}. Here, we adopted the straightforward approach of interpolating the monthly output to pseudo-daily values using a monthly mean-preserving algorithm⁵¹ and then aggregating the daily values back into monthly values using the month-length values appropriate for a particular time as determined by an algorithm⁴⁶ for calculating the celestial longitude or position of Earth along its orbit.

The TraCE simulation is of relatively coarse spatial resolution (model grid of approximately 3.75°) relative to the spatial scale of variability reflected by the pollen data. Consequently, we interpolated the simulated data onto a 0.5° high-resolution grid that is of a fine enough resolution to reflect the spatial variability in the pollen data. The interpolation was done by first calculating 'anomalies' as the differences between individual monthly values of TREFHT at each grid point and a present-day base-period long-term mean for 1961–1990. (The TraCE simulation extends until only 1989, so for 1990 we repeated the 1989 data to calculate the 1961–1990 long-term mean.) These anomalies were then applied to 0.5° averages of the CRU CL 2.0 1961–1990 gridded 'modern' temperature (tmp) data³⁴. This 'apply the anomalies' approach is still the standard approach used when comparing palaeoclimatic simulations with reconstructions⁵², and it reduces bias in the climate-model simulations.

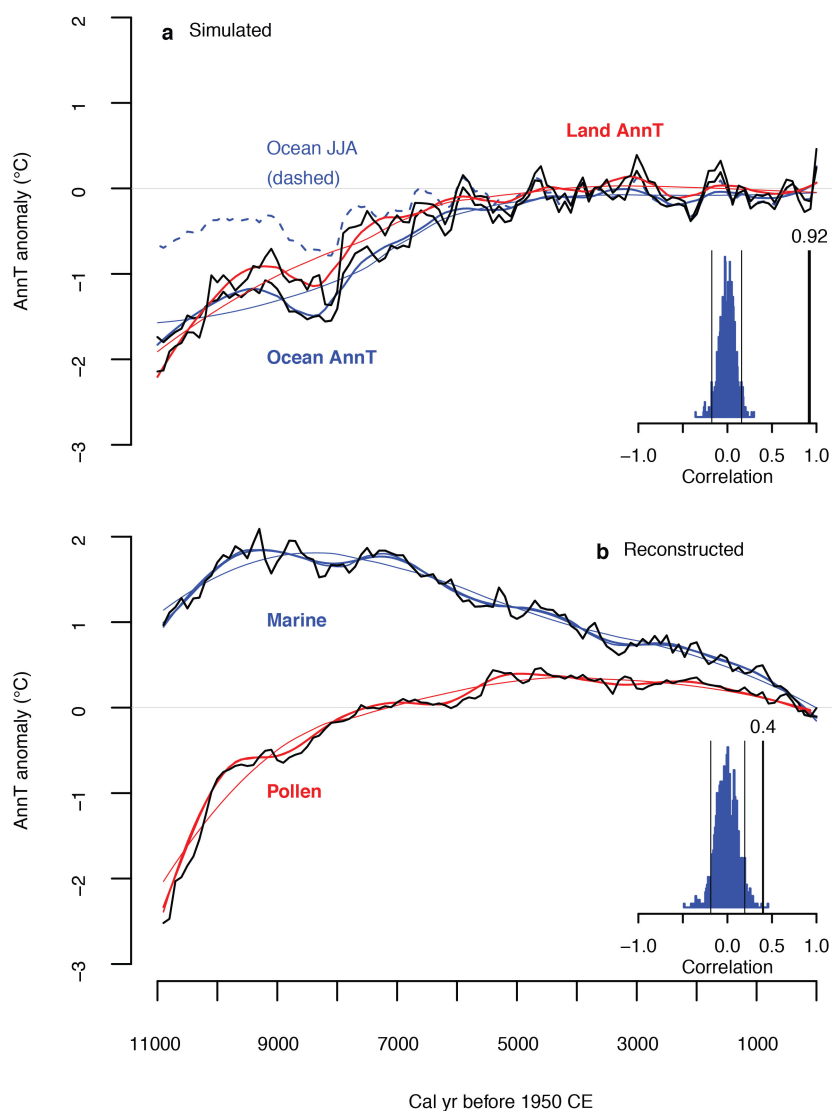
Annual values of AnnT, GDD5 and MTCO were calculated using the 0.5° data, with the same mean-preserving interpolation method as used for the calendar adjustment applied to calculate GDD5. Area averages of the variables were calculated by averaging the 0.5° grid cells in which each climate reconstruction fell, using the ICE-6G ice and land mask to include only non-ice-covered land grid points (<http://www.atmos.physics.utoronto.ca/~peltier/data.php>)^{53,54}. The averages were summarized by fitting a loess curve using a fixed window (half) width of 100 years, the tricube weight function and one 'robustness' iteration⁵⁵.

Code availability. The code used to create and analyse the data used in this study is available in the NOAA National Centers for Environmental Information Paleoclimatology Database (<https://www.ncdc.noaa.gov/paleo/study/22992>).

Data availability. The pollen reconstructions and simulation data that support the findings of this study are available in the NOAA National Centers for Environmental Information Paleoclimatology Database (<https://www.ncdc.noaa.gov/paleo/study/22992>). TraCE-21ka simulations are available at <https://www.earthsystemgrid.org/dataset/ucar.cgd.cesm3.trace.html>.

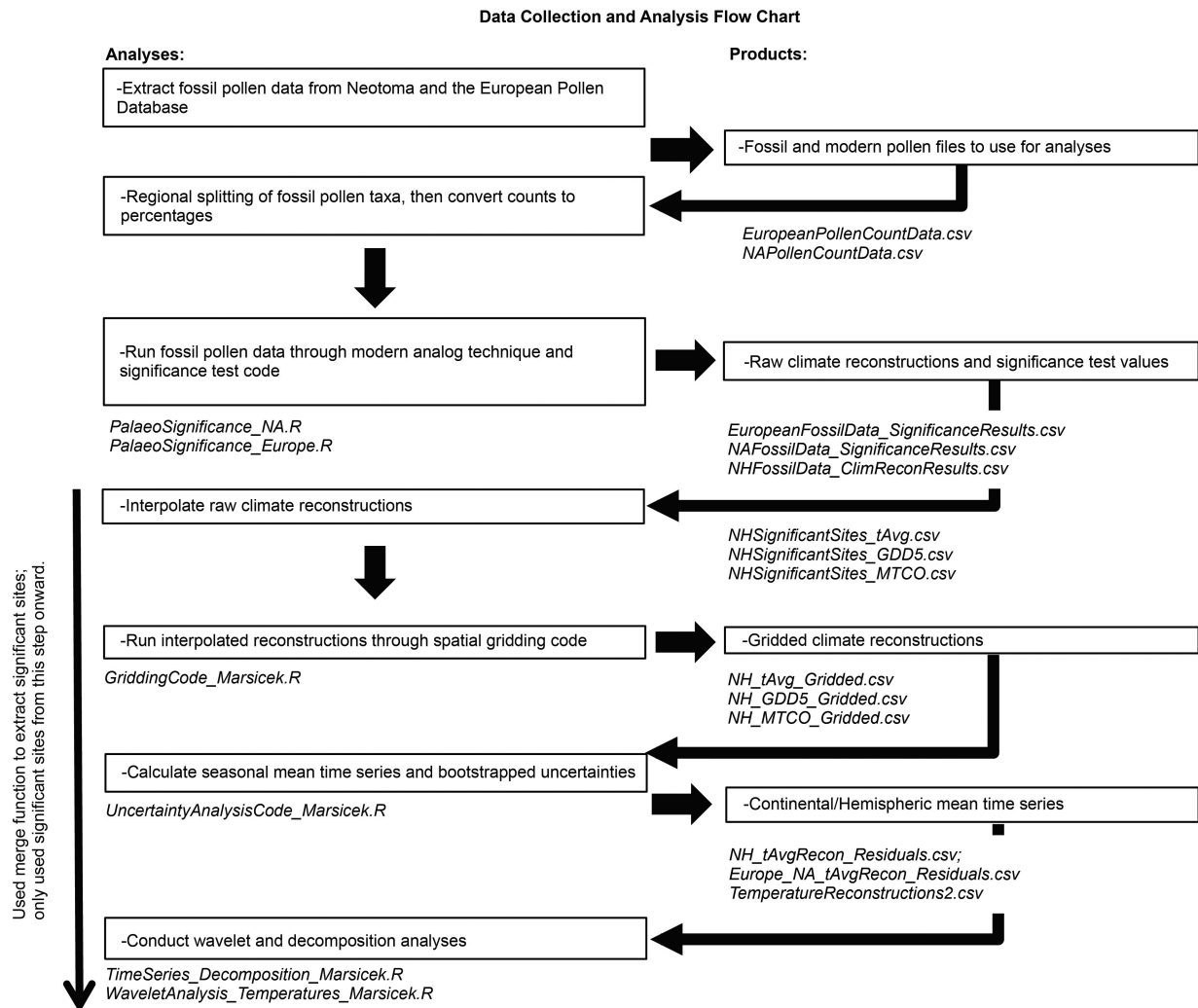
22. Fyfe, R. M. *et al.* The European Pollen Database: past efforts and current activities. *Veg. Hist. Archaeobot.* **18**, 417–424 (2009).
23. Goring, S. *et al.* neotoma: a programmatic interface to the Neotoma Paleocological Database. *Open Quat.* **1**, <http://dx.doi.org/10.5334/oq.ab> (2015).
24. Williams, J. W., Shuman, B., Webb, T., Bartlein, P. J. & Leduc, P. L. Late Quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecol. Monogr.* **74**, 309–334 (2004).
25. Grimm, E. C., Maher, L. J., Jr & Nelson, D. M. The magnitude of error in conventional bulk-sediment radiocarbon dates from central North America. *Quat. Res.* **72**, 301–308 (2009).
26. Blois, J. L., Williams, J. W., Grimm, E. C., Jackson, S. T. & Graham, R. W. A methodological framework for assessing and reducing temporal uncertainty in paleovegetation mapping from late-Quaternary pollen records. *Quat. Sci. Rev.* **30**, 1926–1939 (2011).
27. Lotter, A. F. *et al.* Younger Dryas and Allerød summer temperatures at Gerzensee (Switzerland) inferred from fossil pollen and cladoceran assemblages. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **159**, 349–361 (2000).
28. Marsicek, J. P., Shuman, B., Brewer, S., Foster, D. R. & Oswald, W. W. Moisture and temperature changes associated with the mid-Holocene Tsuga decline in the northeastern United States. *Quat. Sci. Rev.* **80**, 129–142 (2013).
29. Williams, J. W. & Shuman, B. Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset. *Quat. Sci. Rev.* **27**, 669–687 (2008).
30. Whitmore, J. *et al.* An updated modern pollen-climate-vegetation dataset for North America. *Quat. Sci. Rev.* **24**, 1828–1848 (2005).
31. Overpeck, J. T., Webb, T., III & Prentice, I. C. Quantitative interpretation of fossil pollen spectra: dissimilarity coefficients and the method of modern analogs. *Quat. Res.* **23**, 87–108 (1985).
32. Davis, B. A. S. *et al.* The European Modern Pollen Database (EMPD) project. *Veg. Hist. Archaeobot.* **22**, 521–530 (2013).

33. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
34. New, M., Lister, D., Hulme, M. & Makin, I. A high-resolution data set of surface climate over global land areas. *Clim. Res.* **21**, 1–25 (2002).
35. Cheddadi, R., Yu, G., Guiot, J., Harrison, S. P. & Prentice, I. C. The climate of Europe 6000 years ago. *Clim. Dyn.* **13**, 1–9 (1996).
36. Telford, R. & Trachsel, M. *palaeoSigs*: significance tests for palaeoenvironmental reconstructions, R package version 1.1-3 (2015).
37. Juggins, S. *rioja*: analysis of quaternary science data, R package version 0.9-15 (2016).
38. Jackson, S. T. & Williams, J. W. Modern analogs in Quaternary paleoecology: here today, gone yesterday, gone tomorrow? *Annu. Rev. Earth Planet. Sci.* **32**, 495–537 (2004).
39. Telford, R. J. & Birks, H. J. B. A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages. *Quat. Sci. Rev.* **30**, 1272–1278 (2011).
40. Marlon, J. R. *et al.* Climate and human influences on global biomass burning over the past two millennia. *Nat. Geosci.* **1**, 697–702 (2008).
41. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829–836 (1979).
42. Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
43. Gouhier, T. C., Grinstead, A. & Simko, V. *biwavelet*: conduct univariate and bivariate wavelet analyses, R package version 0.20.10 (2016).
44. He, F. *Simulating Transient Climate Evolution of the Last Deglaciation with CCSM3*. PhD Thesis, Univ. Wisconsin-Madison (2011).
45. Joussaume, S. & Braconnot, P. Sensitivity of paleoclimate simulation results to season definitions. *J. Geophys. Res. Atmos.* **102**, 1943–1956 (1997).
46. Kutzbach, J. E. & Gallimore, R. G. Sensitivity of a coupled atmosphere/mixed layer ocean model to changes in orbital forcing at 9000 years B.P. *J. Geophys. Res.* **93**, 803–821 (1988).
47. Bartlein, P. J. & Shafer, S. L. The impact of the “calendar effect” and pseudo-daily interpolation algorithms on paleoclimatic data-model comparisons. In *AGU Fall Meeting Abstracts* abstr. PP31C-2296 (American Geophysical Union, 2016).
48. Timm, O., Timmermann, A., Abe-Ouchi, A., Saito, F. & Segawa, T. On the definition of seasons in paleoclimate simulations with orbital forcing. *Paleoceanography* **23**, PA2221 (2008).
49. Pollard, D. & Reusch, D. B. A calendar conversion method for monthly mean paleoclimate model output with orbital forcing. *J. Geophys. Res. Atmos.* **107**, 4615 (2002).
50. Chen, G.-S., Kutzbach, J. E., Gallimore, R. & Liu, Z. Calendar effect on phase study in paleoclimate transient simulation with orbital forcing. *Clim. Dyn.* **37**, 1949–1960 (2011).
51. Epstein, E. S. On obtaining daily climatological values from monthly means. *J. Clim.* **4**, 365–368 (1991).
52. Harrison, S. P. *et al.* Intercomparison of simulated global vegetation distributions in response to 6 kyr BP orbital forcing. *J. Clim.* **11**, 2721–2742 (1998).
53. Argus, D. F., Peltier, W. R., Drummond, R. & Moore, A. W. The Antarctica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories. *Geophys. J. Int.* **198**, 537–563 (2014).
54. Peltier, W. R., Argus, D. F. & Drummond, R. Space geodesy constrains ice age terminal deglaciation: the global ICE-6G_C (VM5a) model. *J. Geophys. Res. Solid Earth* **120**, 450–487 (2015).
55. Cleveland, W. S. & Devlin, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596–610 (1988).

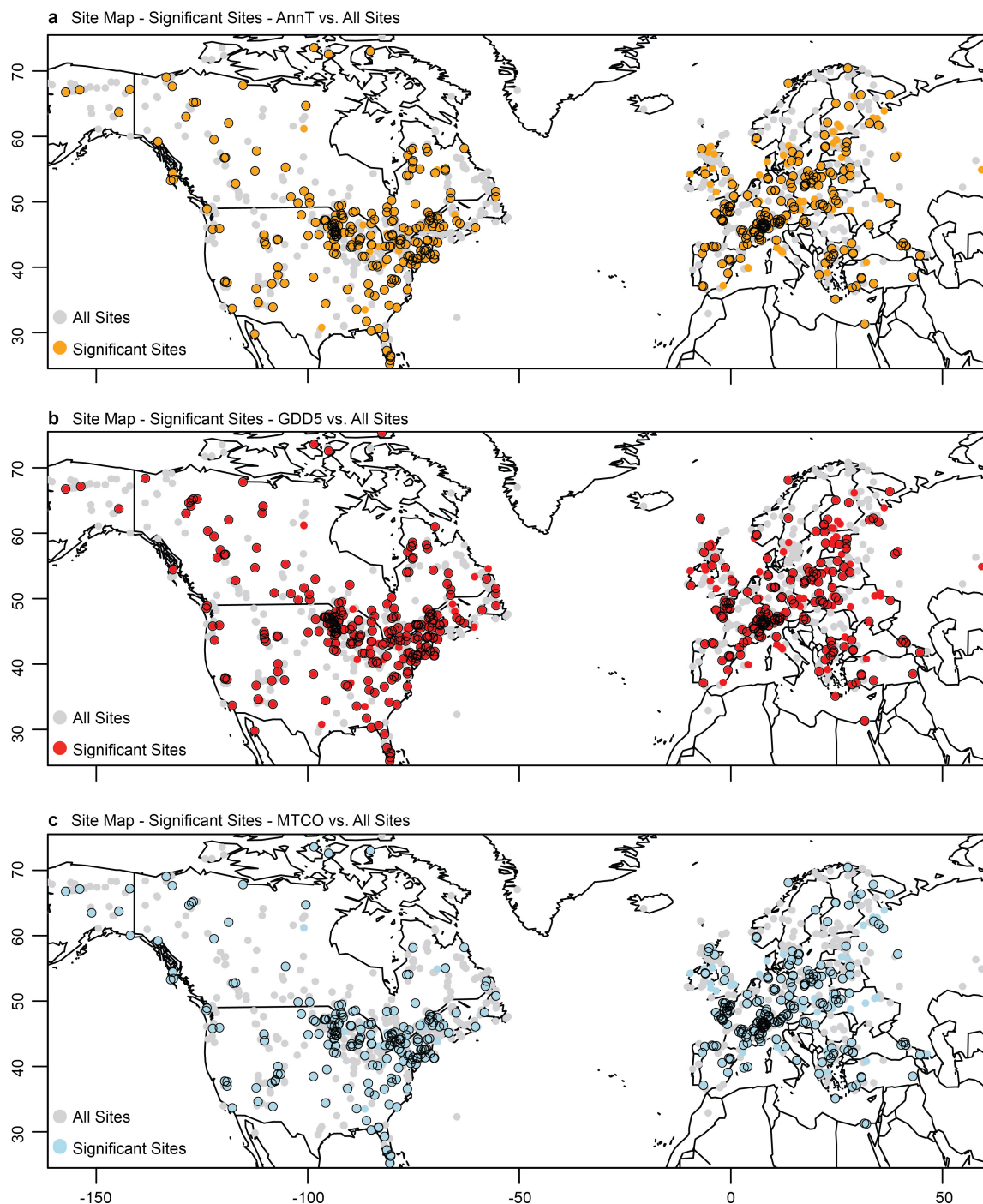


Extended Data Figure 1 | Time series of calendar-adjusted CCSM3-simulated continental and marine AnnTs and pollen-inferred reconstructions of AnnT highlight different long-term trends from the synthesis of marine and coastal temperature reconstructions. **a**, Simulated AnnT anomalies are shown for land (black line with solid red smooth lines), along with the simulated boreal summer (JJA) air-temperature anomalies for ocean grid cells (dashed blue) over the past 11,000 years. Smooth red and blue solid lines indicate the long-term trends in the simulation (6,500-yr loess fits) for comparison with additional millennial-scale variability (captured by 2,500-yr loess fits (thin solid lines) and

GAMMs that account for temporal autocorrelation (thick solid lines)). **b**, Same as **a**, but for the pollen-inferred (red) and marine synthesis (blue) temperature reconstructions. The histograms in the insets show simulated random correlations of simulated or reconstructed AnnT over land with 1,000 random series with the same autoregressive characteristics as the de-trended ocean simulation in Fig. 4a and the de-trended marine reconstruction in Fig. 4b. The thick vertical line represents the correlations between the time series and the thin vertical lines represent the 95% range of the random correlations. All anomalies are relative to the mean of AD 1450–1950. Cal yr before 1950 CE, calibrated years before 1950 of the Common Era.

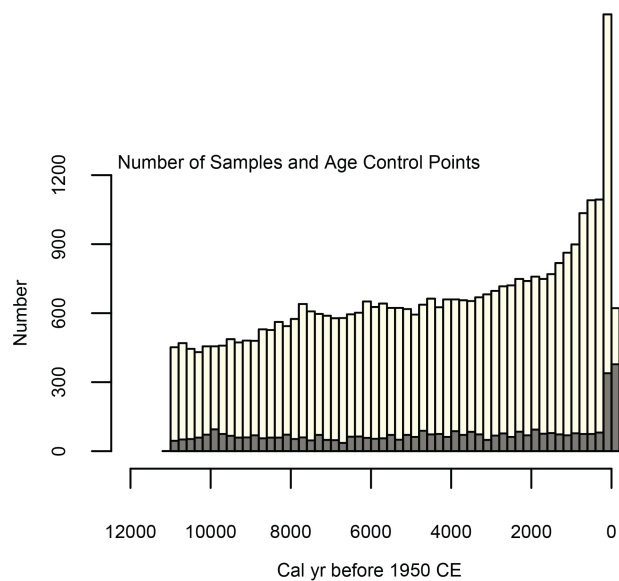


Extended Data Figure 2 | Flow chart for data acquisition, data analyses and data products. Data files and R code used in the reconstruction process can be found at <https://www.ncdc.noaa.gov/paleo/study/22992>.

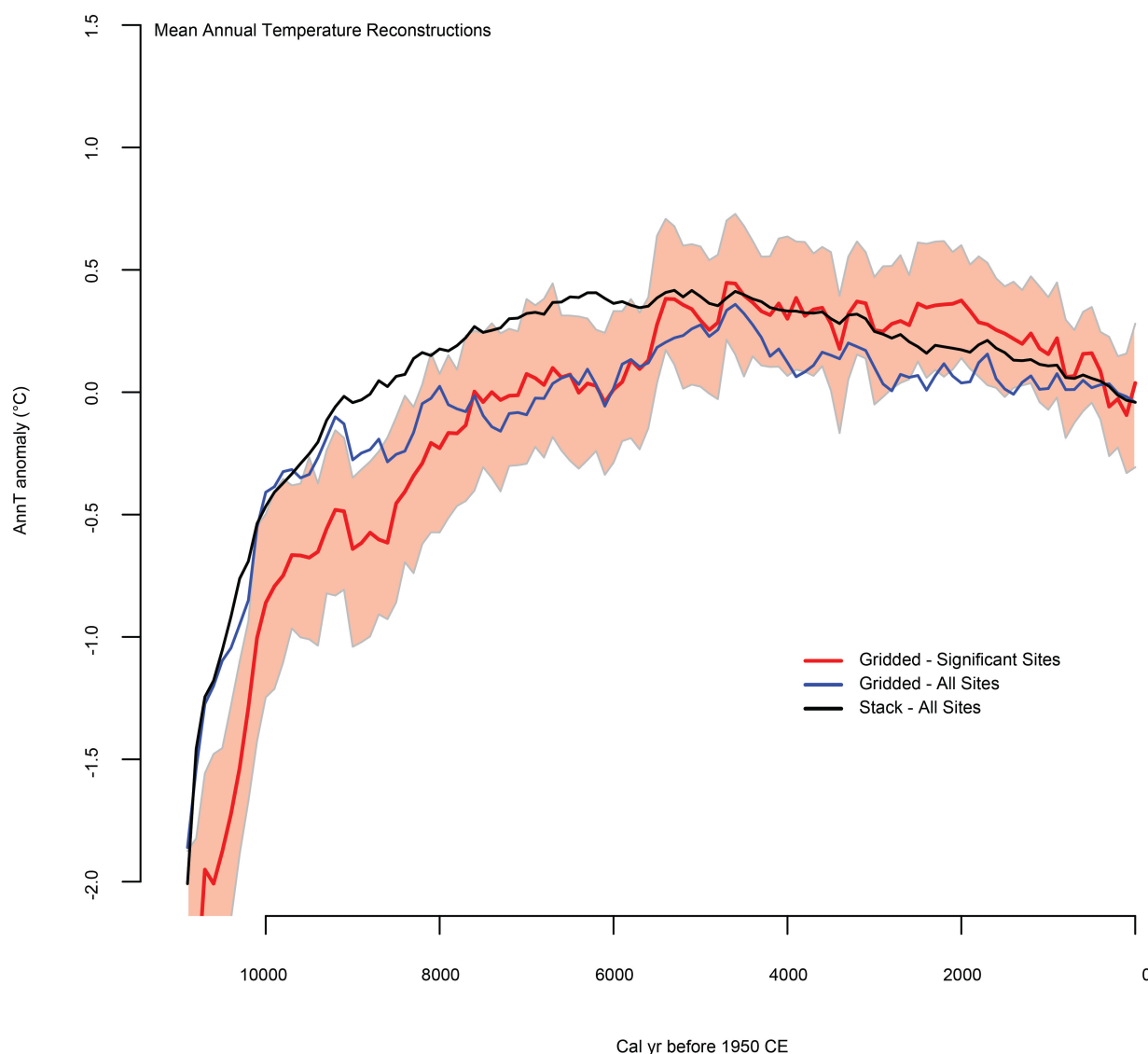


Extended Data Figure 3 | Site maps indicating the location of all and significant pollen-inferred temperature reconstructions. a–c, Maps showing the locations of significant reconstructions (coloured symbols) of AnnT (a), GDD5 (b) and MTCO (c). Grey symbols indicate additional

(not significant) pollen records. Circles with a black outline indicate sites with data within the 0.5–0 kyr BP base period used in calculating temperature anomalies.

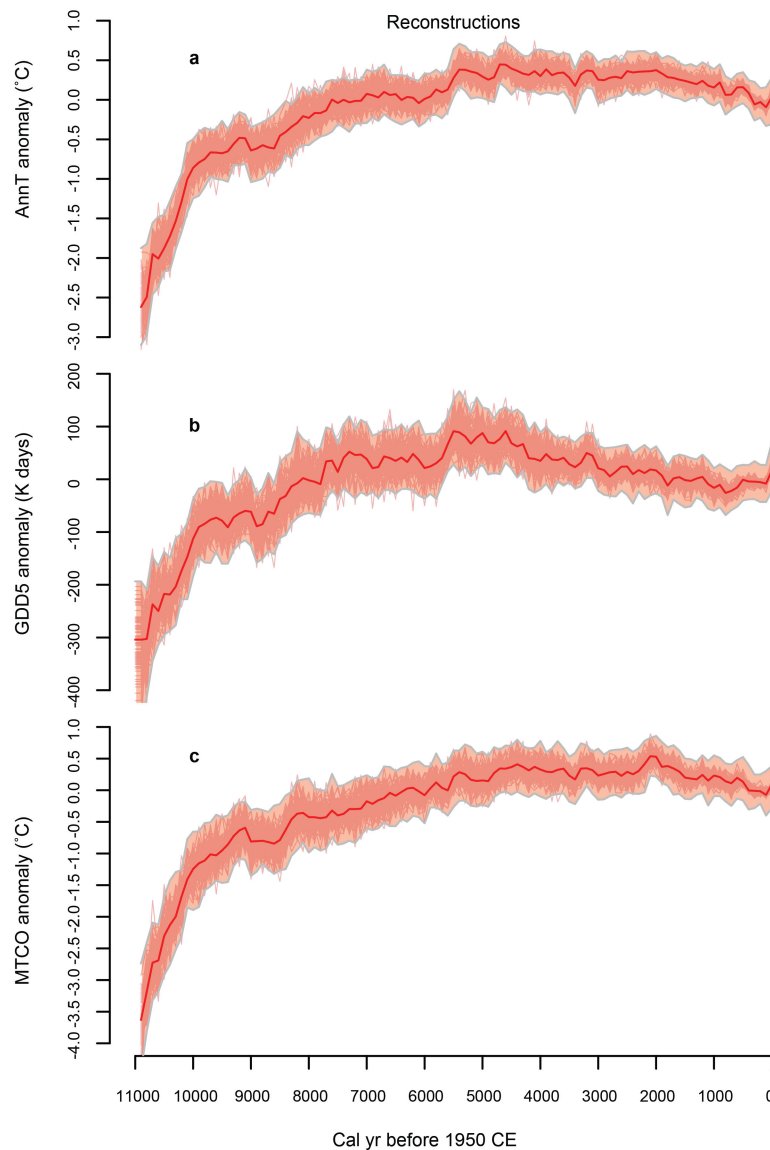


Extended Data Figure 4 | Number of pollen sample and age control points over the Holocene. Number of pollen samples (lightly shaded bars) and dates (dark-shaded bars) remaining after the extraction of significant records.



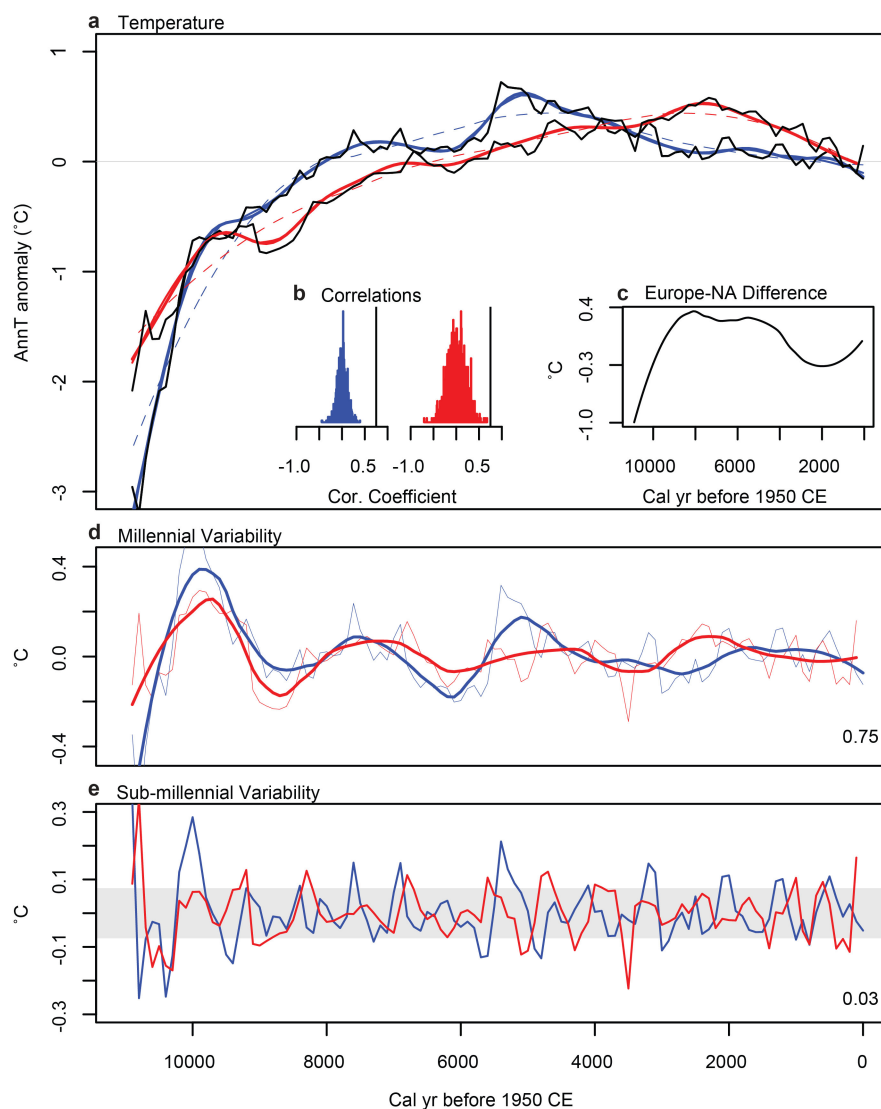
Extended Data Figure 5 | Comparison of the mean of the gridded, significance-screened reconstructions of AnnT with mean reconstructions based on gridding or simply averaging (stacking) all AnnT reconstructions without screening. The AnnT reconstructions for all sites include even the records (grey symbols in Extended Data Fig. 3a) that were not significantly different from random variables, and were produced either by averaging (black line) or first gridding the reconstructions (blue line). These mean time series compare with the mean of the gridded significant reconstructions (coloured symbols in

Extended Data Fig. 3a), which we also use in Figs 1 and 2 (dark red line with uncertainty band). The dark red line represents the median of the 100 individual iterations of the mean in which 50% of grid cells were randomly removed with replacement before the mean was calculated. The underlying red uncertainty band denotes the 2.5%–97.5% quantiles of the 100 individual iterations with noise (based on the RMSE of AnnT) added to each grid cell per century before calculating the mean. All anomalies are relative to the mean of AD 1450–1950.



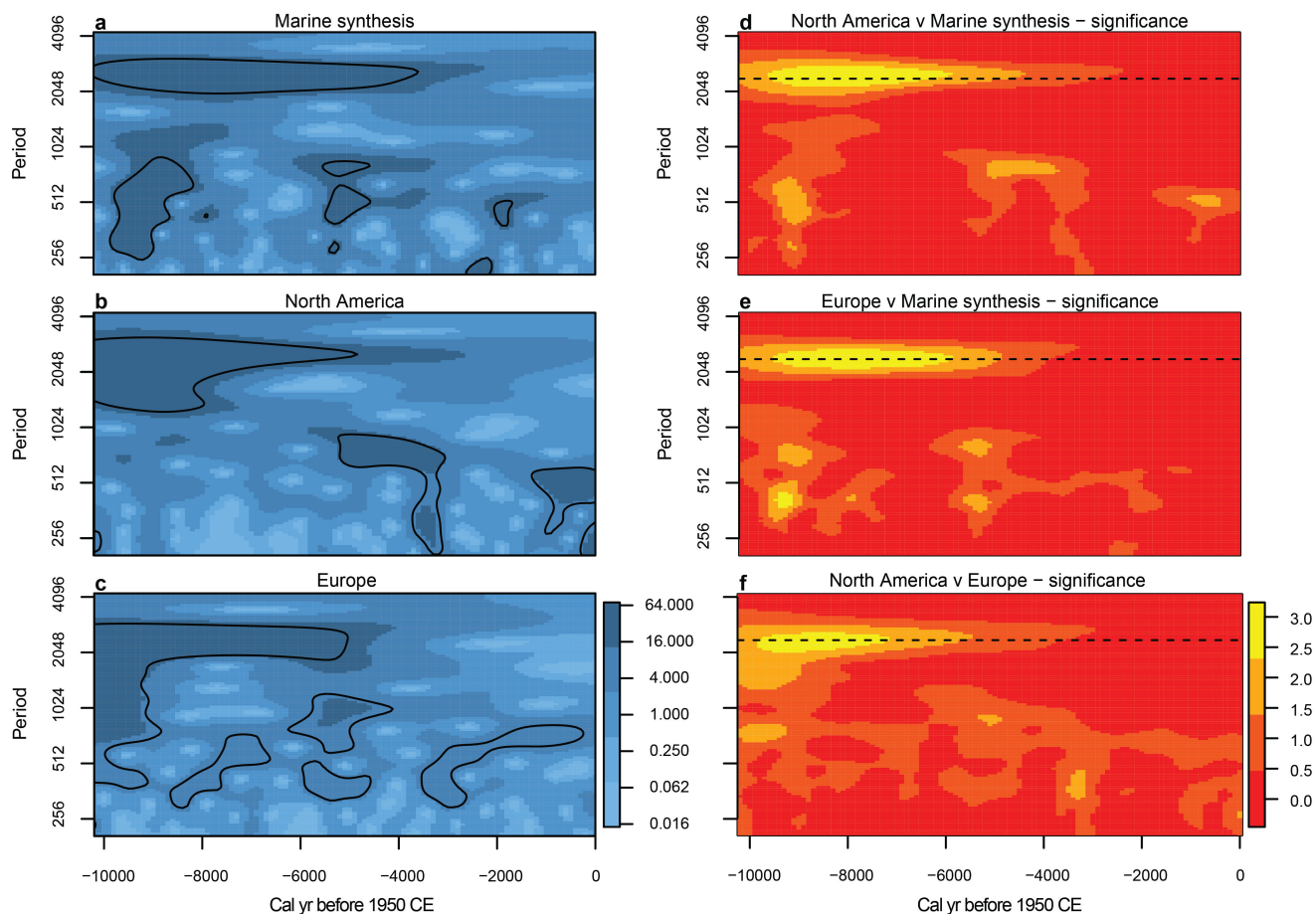
Extended Data Figure 6 | Time series of mean reconstructed temperatures, based on gridding the significant site-level reconstructions, show long-term warming and millennial-scale changes. a–c, AnnT (a), GDD5 (b) and MTCO (c) anomalies reconstructed from 415–453 fossil pollen sites across North America and Europe. Light red lines indicate 100 individual iterations of the reconstruction of each variable reconstruction over the past 11 kyr BP

in which 50% of grid cells were randomly removed with replacement, and the dark red lines represent the median of those iterations. The underlying red uncertainty bands denote the 2.5%–97.5% range of 100 additional iterations that degraded the reconstructions for each gridded cell by adding random noise to each time step based on the RMSE of the modern calibration (see Methods). All anomalies are relative to the mean of AD 1450–1950.



Extended Data Figure 7 | Diagnosing millennial-scale variability from North American and European pollen-inferred AnnTs. **a**, AnnT anomalies for North America (black line with red smooth lines) and Europe (black line with blue smooth lines) over the past 11 kyr BP. Smooth blue and red dashed lines indicate a 6,500-yr loess fit. Solid blue and red lines indicate a 2,500-yr loess fit (thin lines) and a GAMM fit (thick lines) to the North American and European temperature reconstructions after accounting for temporally autocorrelated residuals. **b**, Histograms show correlations of the de-trended North American (red) and European (blue) AnnT reconstructions with 1,000 randomly simulated time series with

the same autoregressive characteristics as the de-trended reconstructions (thin lines in **d**). Vertical black lines indicate the correlation between the two de-trended reconstructions ($r = 0.75$). **c**, Difference between the 6,500-yr loess fits to the North American and European reconstructions. **d**, Difference between the European and North American means (thin lines) and GAMMs (bold lines) after de-trending by subtracting the 6,500-yr loess fits. **e**, Same as in **d**, but removing a 2,500-yr loess fit. The grey bar indicates departures of one standard deviation from zero. All anomalies are relative to the mean of AD 1450–1950.



Extended Data Figure 8 | Wavelet analysis highlights coherent millennial-scale variability. **a–c**, Wavelet decomposition of the marine synthesis (**a**), the North American AnnT reconstruction (**b**) and the European AnnT reconstruction (**c**); the colour scale represents the power spectrum (darker colours represent more power). **d–f**, Cross-wavelet significance is also shown for North America versus the marine synthesis (**d**),

Europe versus the marine synthesis (**e**) and North America versus Europe (**f**); the colour scale indicates whether there is a significant relationship between the cross-wavelets of the datasets (red, no significant relationship; yellow, significant relationship at the 0.9 significance level). The period, in years, is shown on the y axes.

Extended Data Table 1 | Modern calibration experiment results for AnnT

Experiment	Grid average	No. Analogs	R ²	RMSE	Percent loss of skill	N
2-degree grid means, NA-E	Yes	7	0.945	2.19	4.8	820
2-degree grid means, NA-E	Yes	1	0.942	2.23	6.7	820
Leave-one-out point-to-point comparison, NA-E	No	7	0.924	2.09	0.0	8564
Leave-one-out point-to-point comparison, NA-E	No	1	0.911	2.29	9.6	8564
1 site per 2-degree grid, NA-E	No	7	0.911	2.78	33.0	820
1 site per 2-degree grid, NA-E	No	1	0.9	2.91	39.2	820

Loss of skill is the difference in RMSE between an experiment and the experiment with the lowest RMSE, divided by the experiment with the lowest RMSE (expressed as a percentage).

Extended Data Table 2 | Sites used in marine synthesis

Marcott Unique ID	Longitude	Latitude	Used in Marine Synthesis
1	34.7	27.7	No
2	-124.9	41.7	Yes
4	-18	67	Yes
6	-2.6	36.1	Yes
7	-7.1	39.4	No
8	7.6	67	Yes
9	-24.1	61.4	No
10	34.1	31.7	Yes
11	-143.6	61.4	Yes
13	7.6	67	No
14	7.1	57.7	No
15	8.7	57.8	Yes
16	17.7	36.7	Yes
18	18.6	40.9	No
30	-7.1	36.4	Yes
32	-138.4	68.4	Yes
34	19.2	68.4	Yes
35	18.7	68.4	Yes
42	-26	58.8	Yes
51	-9.5	38.6	Yes
53	-55	43	Yes
54	-63	44	Yes
55	-74.6	36.9	Yes
56	14	75	No
58	13.7	58.6	Yes
59	22.1	68.7	Yes
65	-17.8	62.1	Yes
73	-87.1	29	No

Sites are from ref. 3.

Early Middle Palaeolithic culture in India around 385–172 ka reframes Out of Africa models

Kumar Akhilesh¹, Shanti Pappu¹, Haresh M. Rajapara^{2,3}, Yanni Gunnell⁴, Anil D. Shukla⁵ & Ashok K. Singhvi³

Luminescence dating at the stratified prehistoric site of Attirampakkam, India, has shown that processes signifying the end of the Acheulian culture and the emergence of a Middle Palaeolithic culture occurred at 385 ± 64 thousand years ago (ka), much earlier than conventionally presumed for South Asia¹. The Middle Palaeolithic continued at Attirampakkam until 172 ± 41 ka. Chronologies of Middle Palaeolithic technologies in regions distant from Africa and Europe are crucial for testing theories about the origins and early evolution of these cultures, and for understanding their association with modern humans or archaic hominins, their links with preceding Acheulian cultures and the spread of Levallois lithic technologies^{2–20}. The geographic location of India and its rich Middle Palaeolithic record are ideally suited to addressing these issues, but progress has been limited by the paucity of excavated sites and hominin fossils as well as by geochronological constraints^{1,8}. At Attirampakkam, the gradual disuse of bifaces, the predominance of small tools, the appearance of distinctive and diverse Levallois flake and point strategies, and the blade component all highlight a notable shift away from the preceding Acheulian large-flake technologies⁹. These findings document a process of substantial behavioural change that occurred in India at 385 ± 64 ka and establish its contemporaneity with similar processes recorded in Africa and Europe^{2–8,10–13}. This suggests complex interactions between local developments and ongoing global transformations. Together, these observations call for a re-evaluation of models that restrict the origins of Indian Middle Palaeolithic culture to the incidence of modern human dispersals after approximately 125 ka^{19,21}.

The end of the Lower Palaeolithic Acheulian culture and beginnings of the Middle Palaeolithic, or Middle Stone Age, involved processes that marked substantial changes in hominin behaviour. The legacy of these changes, placed at approximately 300–200 ka^{2–8}, is expressed primarily through technological transformations that involve a gradual decline in Acheulian large flake and core tools⁹, including bifaces; a proliferation and diversity of Levallois flake- and point-reduction strategies; and the evolution of blade technologies^{3–7,10,11}. The behavioural processes that underpinned the transition from the Acheulian to the early Middle Palaeolithic or Middle Stone Age were variable and complex through space and time. This is evident at several Middle Palaeolithic and Middle Stone Age sites from the continuation of biface production—characteristic of Acheulian cultures—in small numbers amidst diverse Levallois- and blade-reduction sequences, and from the Acheulian roots of the Levallois concept^{8,10,13–17} (see Supplementary Information). The co-occurrence of Middle Palaeolithic or Middle Stone Age artefact sequences with not only modern humans² but also other archaic species—with which modern humans could potentially interact—complicates investigations considerably^{7,8,14} (see Supplementary Information).

Despite the presence of numerous Middle Palaeolithic sites in South Asia, the age and origin of this cultural phase remain poorly

documented^{8,18} (see Supplementary Information). Important features of the Middle Palaeolithic in India include the continuation of bifaces (albeit occurring less frequently or smaller in size than their Acheulian analogues); a predominance of small flake tools; the presence of Levallois and blade technologies and occasional points; and in some regions, depending on availability, an increased preference for fine-grained cryptocrystalline raw materials^{8,18} (see Supplementary Information). Radiometric ages have so far placed Indian Middle Palaeolithic cultures at approximately 140–46 ka^{1,20}, potentially overlapping with a possible Late Acheulian occurrence at approximately 140–120 ka¹⁹. Regional variants and evolutionary trajectories of the Indian Middle Palaeolithic, and its association with modern humans or other archaic species and with the origins of Levallois technology, continue to be debated^{8,21}. Patterns of hominin dispersals inferred from correlations between genetic, fossil and archaeological records are likewise unclear⁸. One theory²² links the Middle Palaeolithic in India with modern human dispersals out of Africa during and after Marine Isotope Stage 5 (130–80 ka), with populations surviving the catastrophic Toba volcanic eruptions at around 74 ka, whereas a contrasting theory^{20,23} associates the Indian Middle Palaeolithic with coexisting archaic species and advocates that the arrival of modern humans—ushering in microlithic blade assemblages and other cultural features—did not occur before Marine Isotope Stage 4 or 3 (71–57 ka). These gaps in our understanding of cultural transformations in South Asia arise from the scarcity of radiometric ages at excavated sites and of hominin fossils.

Here we present chronological and archaeological evidence from Attirampakkam (ATM), a Lower and Middle Palaeolithic site situated on the banks of a tributary stream of the Kortallaiyar River²⁴ (Fig. 1, Extended Data Fig. 1). Excavations to depths of between 4 and 9 m in different trenches have revealed an alluvial sequence deposited by a small stream transporting a sediment load derived from shale, sandstone and laterite outcrops. From the base upwards, layers 8 to 6 are clay-rich and contain exclusively Early Acheulian assemblages (dating to approximately 1.7–1.07 million years ago (Ma))²⁴; the overlying layers 5 to 1 contain the Middle Palaeolithic assemblages and form a sequence of clay-rich silt alternating with ferruginous gravel (Fig. 2, Extended Data Fig. 2). The mineral magnetic record²⁵ indicates a seasonally dry tropical climate that was wetter during the deposition of layers 4 and 3, which are low-energy overbank silt deposits, and drier during the deposition of layers 5 and 2, which are gravel beds, with aridity persisting through layer 1 (see Supplementary Information).

Our description of the composition of the Middle Palaeolithic assemblage is based on the contents of three adjoining trenches (T7A, T7B and T7C) and involves the systematic analysis of 7,261 artefacts excavated from trench T7A (Figs 3, 4, Extended Data Figs 2–8, Supplementary Table 1). Like their Acheulian predecessors²⁴, Middle Palaeolithic populations used locally available quartzite for making tools: other siliceous rock sources are absent in the region²⁶.

¹Sharma Centre for Heritage Education, 28 1st Main Road, C.I.T. Colony, Mylapore, Chennai 600004, Tamil Nadu, India. ²Department of Physics, Electronics and Space Science, Gujarat University, Navrangpura, Ahmedabad 380009, India. ³AMOPH Division, Physical Research Laboratory, Navrangpura, Ahmedabad 380009, India. ⁴Université de Lyon, Department of Geography, UMR 5600 Environnement Ville Société, 5 Avenue Pierre Mendès-France, F-69696 Bron, France. ⁵Geosciences Division, Physical Research Laboratory, Navrangpura, Ahmedabad 380009, India.

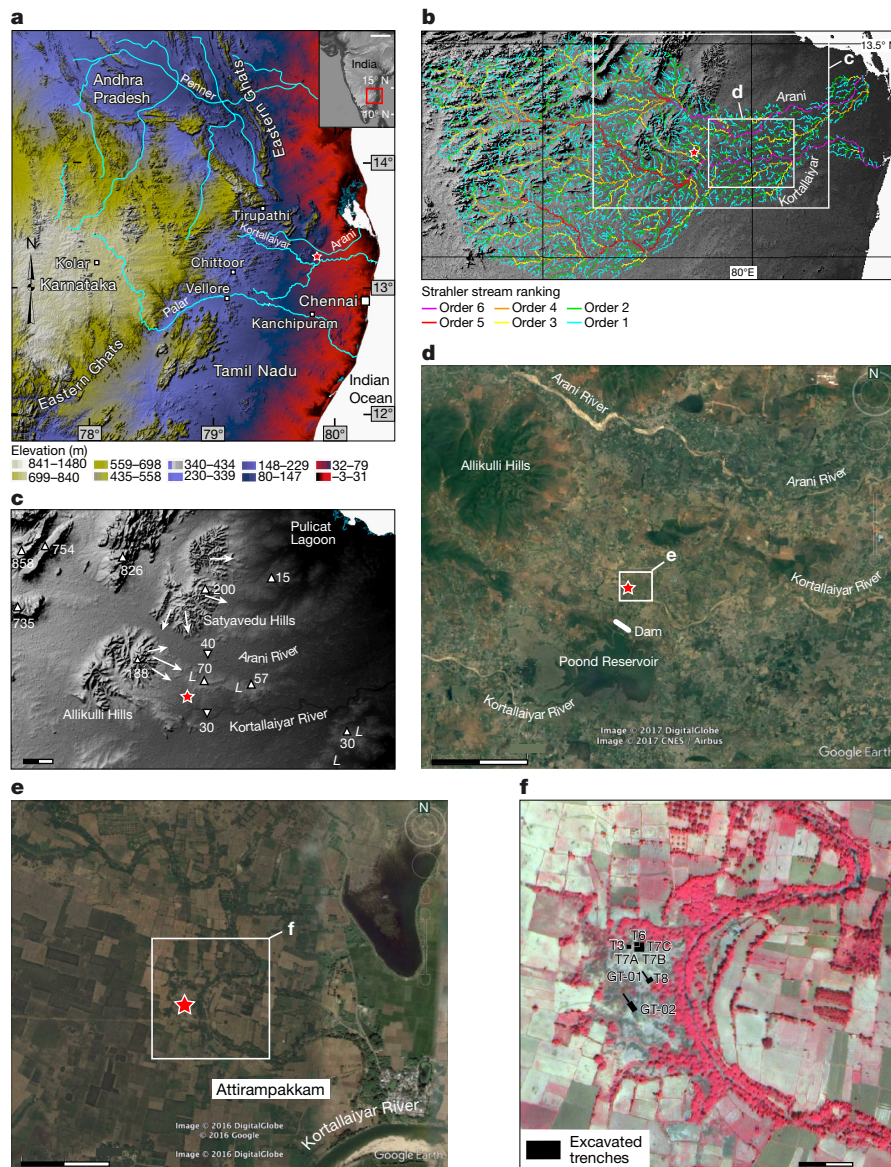


Figure 1 | Location of ATM. In all panels, ATM is indicated by a red star. **a**, Regional setting. Scale bar (inset), 400 km. **b**, Drainage pattern, coded by Strahler stream order. **c**, Regional topography. The Allikulli and Satyavedu Hills are the stumps of Mesozoic alluvial fans. Their quartzite clasts have been redistributed eastward (white arrows) as loose debris by streams and slope processes, and thereby became available for toolmaking. L, laterite. Upward-pointing triangle, spot height of hill (elevation in metres above sea level); downward-pointing triangle, spot height of stream channel (elevation in metres above sea level). Scale bar, 4 km. **d**, Map highlighting

the contrast between the wide ribbons of gravel formed by order-6 braided rivers, and those formed by the smaller order-3 ATM stream. Scale bar, 6 km. **e**, The meandering ATM stream amidst cultivated fields. Scale bar, 420 m. **f**, False-colour satellite image (IKONOS) of the excavations. Reprinted from ref. 26, with permission from Elsevier. Red tone: woody vegetation. Green and yellow: crops or bare soil. Scale bar, 80 m. Image sources: **a–c**, Shuttle Radar Topography Mission 90 m Digital Elevation Data (CGIAR-CSI); **d, e**, Google, CNES/Airbus, DigitalGlobe.

Quartzite occurred locally at the site as pebbles or cobbles in layer 5, but clasts suitable for the production of small flakes were absent in the finer-grained sediments of layers 4 to 1. As a result, quartzite clasts were instead collected from within a radius of 5–10 km from ATM, a landscape that is replete with other Middle Palaeolithic sites²⁶ (Extended Data Fig. 1). Manuports in exotic raw material found at ATM included a tool on silicified wood and an unmodified quartz crystal (Fig. 4s).

Sediment samples from layers 5 to 1 in trench T7A were dated using post-infrared infrared-stimulated luminescence (pIR-IRSL) (Supplementary Tables 2–4, Extended Data Figs 9, 10; see Supplementary Information for methodological details of luminescence dating). The age sequence indicates three main chronological phases (Fig. 2, Supplementary Tables 2–4, Extended Data Figs 9, 10); of these, phases I and II correlate with changes observed in the cultural sequence. The technology associated with phase I (385 ± 64 ka) is confined to

layer 5. A key feature of this technology is the almost complete abandonment of Acheulian large-flake strategies; these earlier strategies produced large cutting tools, including handaxes and cleavers, of greater than 10 cm maximum dimensions (Fig. 3, Extended Data Fig. 3). The sporadic occurrence of bifaces (including diminutive examples) and of a few large flakes suggests the persistence of Acheulian technological skills among the early Middle Palaeolithic hominin groups at ATM. This phenomenon has also been encountered at sites in Africa and parts of Europe^{10,13,15} (see Supplementary Information). A few handaxes that display a preferential flake-removal scar (Extended Data Fig. 3i) provide a hint that the Levallois technique may have possibly been derived from bifacial knapping strategies¹⁷. Small cores, including preferential and recurrent Levallois cores aimed at the production of small flakes, abound in layer 5 (Fig. 3, Extended Data Figs 3, 7b, c, 8) and—along with Levallois points—indicate that proficiency in Middle

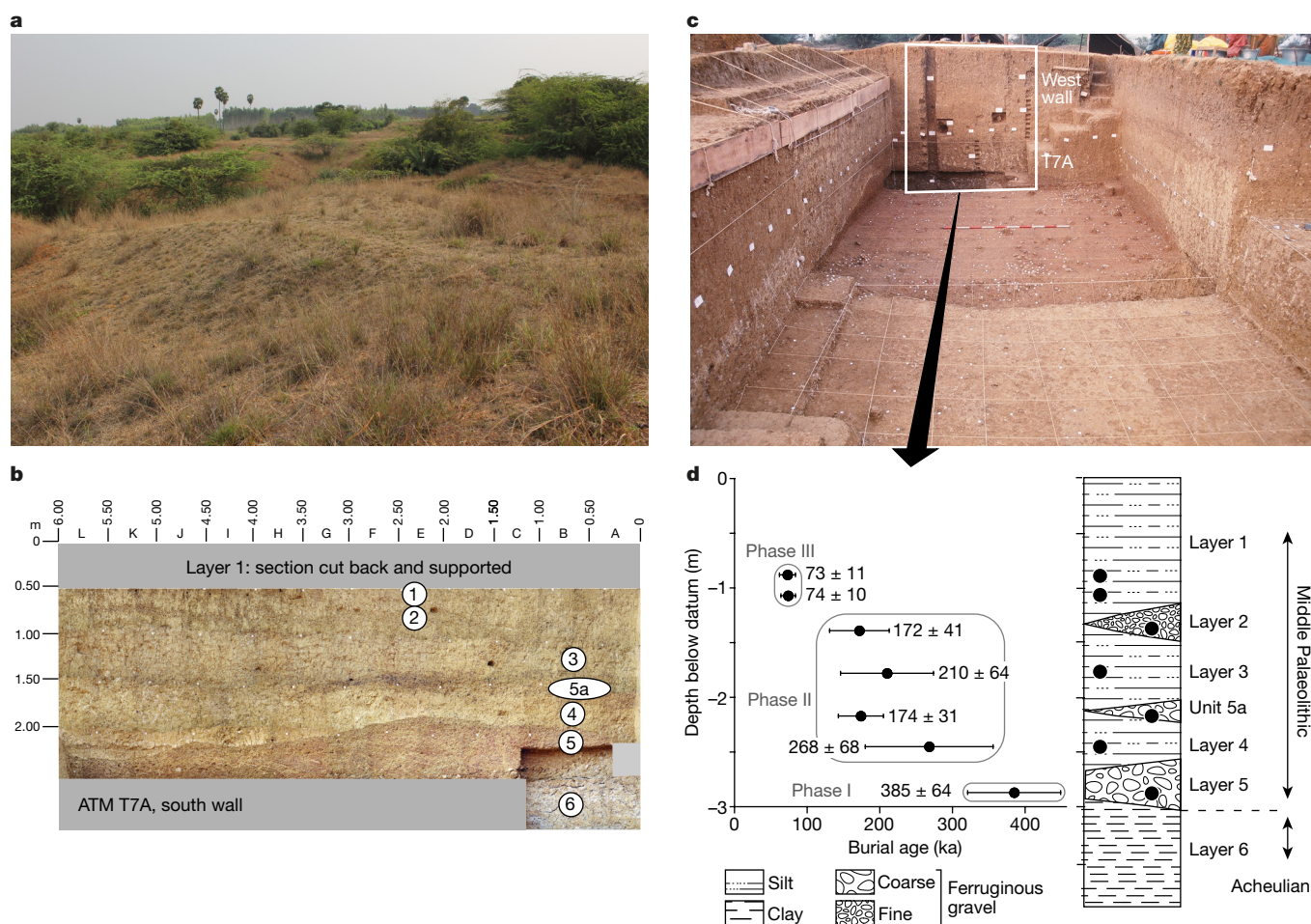


Figure 2 | Stratigraphy, cultural phases and chronology of MP layers at ATM. **a**, General view of the site. **b**, Stratigraphic sequence at ATM showing layers 6 to 1. **c**, **d**, The stratigraphic section on the west wall

Palaeolithic knapping skills was well-established during phase I. These features were entirely absent from the earlier Acheulian assemblages at ATM²⁴. Phase I was also associated with a high frequency of small retouched flake tools, which included scrapers, points on flakes, bifacially flaked points and a tanged point (see Supplementary Information). Incipient blade cores and associated debitage were also present (Extended Data Fig. 6). These features suggest that the behavioural changes that led to the establishment of an early Middle Palaeolithic culture at ATM were occurring during phase I.

The ATM Middle Palaeolithic phase II comprises layers 4 (268 ± 68 ka) and 3 (210 ± 64 ka). In this phase, Levallois strategies for the production of small flakes, blades, points and scrapers were a feature that continued from phase I (Fig. 4, Extended Data Fig. 4). A greater proficiency in blade detachment was evidenced, with the presence of uni- or bi-directional blade removals and associated debitage. In addition, Middle Palaeolithic assemblages in layer 2 (172 ± 41 ka) display up-sequence continuity in terms of the Levallois and blade techniques (Fig. 4, Extended Data Fig. 4).

The low artefact density in layer 1 (see Supplementary Information) precludes any robust definition of a phase III (which, if present, would date to approximately 74 ± 10 ka). However, the sharp drop in artefact densities in this layer suggests a decline in hominin occupation. Despite the absence of ash deposits at ATM, the fact that this sharp decline coincides with dates for the Toba volcanic super-eruption²⁷ could suggest an environmental cause for site abandonment.

The silt-dominated sedimentary sequence at ATM throughout layers 8–1²⁴ suggests that the site was situated in a relatively low-energy

floodplain environment compared to the substantially wider and more energetic neighbouring rivers such as the Kortallaiyar and Arani, which rank much higher in the stream-order hierarchy (Fig. 1).

of trench T7A that was dated by pIR-IRSL. Age clusters define Middle Palaeolithic cultural phases (error bars: 1σ). See Extended Data Fig. 2 for trench details.

Despite the antiquity of the Middle Palaeolithic artefacts from layer 5 (385 ± 64 ka; Fig. 2), their burial age nonetheless suggests that there was a long hiatus between Early Acheulian occupation (1.7–1.07 Ma)²⁴ and the processes that led to the Middle Palaeolithic technological changes. This is inferred from the stratigraphic unconformity that separates layer 6 from layer 5. A Late Acheulian phase has been documented elsewhere in the region and in India more generally (see Supplementary Information and references therein); because this phase is absent at ATM, we infer that the site was temporarily abandoned during the Late Acheulian period —probably as result of local rather than regional causes, whether environmental or otherwise.

Changes in the cultural sequence that are associated with phase I establish that processes that signal the end of the Terminal Acheulian culture and transitions that mark the beginning of the Indian Middle Palaeolithic were occurring between approximately 450 and 320 ka. Notably, phase I includes Marine Isotope Stage 11²⁸, which corresponds to a time period in which the global climate was similar to that during Marine Isotope Stage 5e and the Holocene. The warmer and wetter conditions associated with Marine Isotope Stage 11 would have been conducive to long-distance hominin dispersals, minimally obstructed by greener deserts between Sub-Saharan Africa and South Asia. Phase II spanned a succession of global Middle Pleistocene climatic changes, with little discernible influence on the intensity or continuity of site occupation. The overlap between ages at

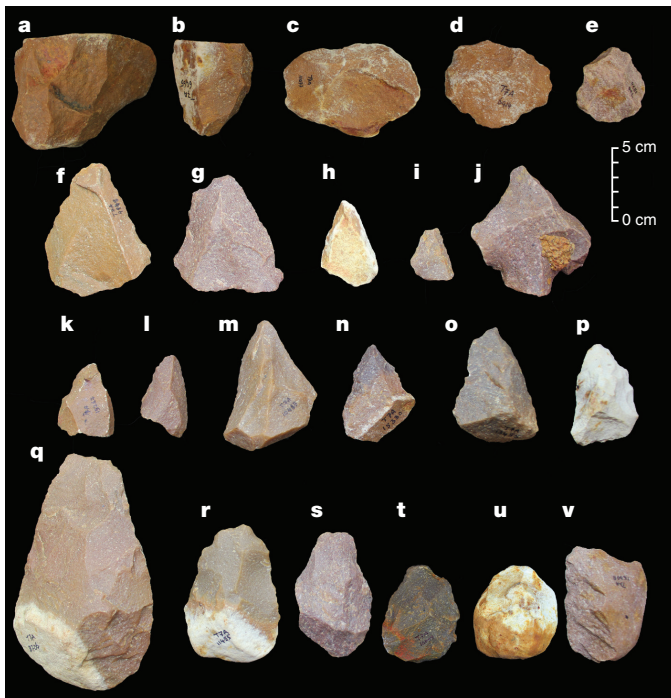


Figure 3 | Representative artefacts from layer 5, trench T7A. a, b, Blade cores (see Extended Data Fig. 6a for details of blade removals in a). c–e, Levallois cores. f–i, Levallois points. j, Tanged point (see Extended Data Fig. 5a for details). k–n, Retouched points. o, p, Bifacially flaked points. q–u, Handaxes (r–u are diminutive). v, Diminutive cleaver.

ATM and the few dated Late Acheulian sites elsewhere in India¹⁹ also suggests that spatial variability among Palaeolithic cultural sequences is larger than previously thought. The assemblage structure at ATM thus adds not only antiquity but also diversity to the mosaic of younger Middle Palaeolithic sites that have previously been described²².

The behavioural transformations that mark the advent of the Indian Middle Palaeolithic at ATM are summarized by the following diagnostic features: the obsolescence of Acheulian large-flake reduction sequences, with a directional shift towards smaller tool components; the adoption and continuance of Levallois recurrent and preferential strategies; a gradual intensification of blade reduction; and an increased use of finer-grained quartzite during phase II than during earlier occupations. A gradual discontinuation of biface use—which becomes definite at ATM after approximately 172 ± 41 ka—has been reported at other Middle Palaeolithic and Middle Stone Age sites worldwide (see Supplementary Information and references therein). Accordingly, and given the well-recognized complexity of cultural transitions^{7,8,10–13}, where bifaces occasionally occur amidst reduction sequences that overwhelmingly suggest new technical preferences and behavioural strategies, it would be inappropriate to use sporadic bifaces as supporting evidence for the persistence of a separate Acheulian culture (see Supplementary Information).

Conclusive correlations between the Middle Palaeolithic assemblages at ATM and a specific hominin species^{2,8,21}—whether modern humans² or archaic hominins^{29,30}—cannot be established because India currently lacks fossil or genetic evidence for this time period other than the Narmada fossil cranium, which could signal the late survival of an archaic species (see Supplementary Information). Evidence of distinct behavioural changes is nonetheless provided by the assemblage structure in the form of new technological strategies, which retain minor components of an archaic nature at around 385 ± 64 ka but depart considerably from Acheulian strategies, and which evolved at ATM for approximately another 200 ka. These processes thus establish the presence of a fully fledged Middle

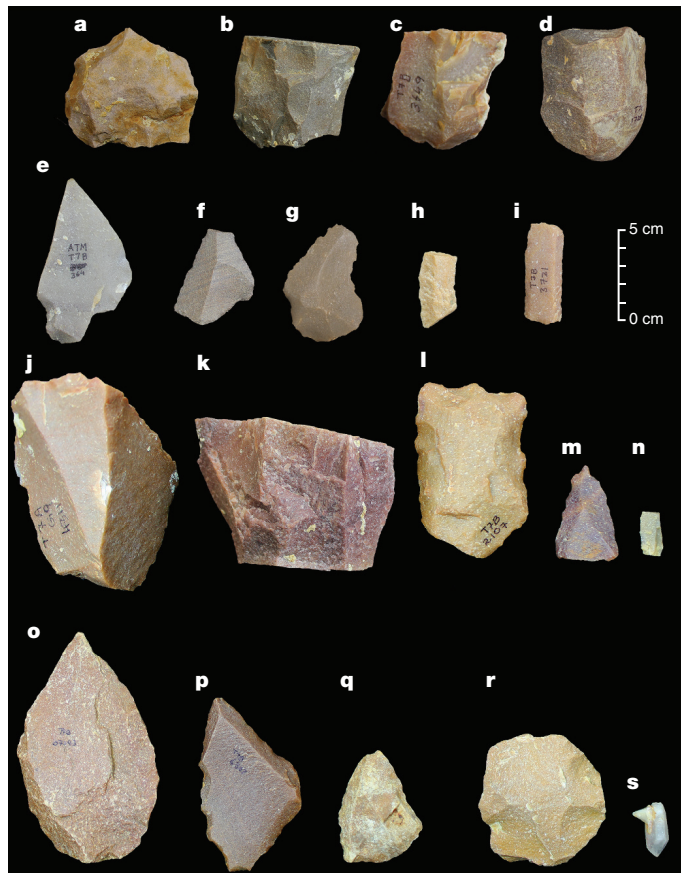


Figure 4 | Representative artefacts from layers 2, 3 and 4 in trenches T7A, T7B and T7C. a–i, Representative artefacts from layer 2: Levallois flake core (a), blade cores (b–d), tanged point (e) (see Extended Data Fig. 5b for enlarged images showing details), Levallois point with broken tip (f), scraper on Levallois flake (g) and blades (h, i). j–n, Representative artefacts from layer 3: blade cores (j, k), diminutive cleaver (l), Levallois point (m) and blade fragment (n). o–s, Representative artefacts from layer 4: point (o), tanged point (p), Levallois point (q), Levallois flake core (r) and quartz crystal manuport (s).

Palaeolithic culture in India at around 385–172 ka, which long pre-dates any previous evidence that suggests Middle Palaeolithic technologies were disseminated out of Africa by modern humans from around 125 ka or later^{1,8,20–23}. The respective parts played by local rather than external influences in the early rise of Middle Palaeolithic culture in India remain uncertain. However, when set in a global context (see Supplementary Information and references therein), the sequence at ATM suggests a succession of population dispersals across South Asia during the Middle Pleistocene, which perhaps involved interactions with other archaic species.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 15 June; accepted 9 December 2017.

1. Singhvi, A. K. *et al.* A ~200 ka record of climatic change and dune activity in the Thar Desert, India. *Quat. Sci. Rev.* **29**, 3095–3105 (2010).
2. Richter, D. *et al.* The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature* **546**, 293–296 (2017).
3. Tryon, C. A. 'Early' Middle Stone Age lithics of the Kapthurin Formation (Kenya). *Curr. Anthropol.* **47**, 367–375 (2006).
4. Tryon, C. A. *The Acheulian to Middle Stone Age Transition: Tephrostratigraphic Context for Archaeological Change in the Kapthurin Formation, Kenya* (Univ. Connecticut Press, 2003).
5. Sahle, Y., Morgan, L. E., Braun, D. R., Atanfu, B. & Hutchings, W. K. Chronological and behavioral contexts of the earliest Middle Stone Age in the Gademotta Formation, Main Ethiopian Rift. *Quat. Int.* **331**, 6–19 (2014).

6. Porat, N. *et al.* New radiometric ages for the Fauresmith industry from Kathu Pan, southern Africa: implications for the Earlier to Middle Stone Age transition. *J. Archaeol. Sci.* **37**, 269–283 (2010).
7. Álvarez-Alonso, D. First Neanderthal settlements in northern Iberia: the Acheulean and the emergence of Mousterian technology in the Cantabrian region. *Quat. Int.* **326–327**, 288–306 (2014).
8. James, H. V. A. & Petraglia, M. D. Modern human origins and the evolution of behavior in the Late Pleistocene record of South Asia. *Curr. Anthropol.* **46**, S3–S27 (2005).
9. Sharon, G. *Acheulian Large Flake Industries: Technology, Chronology, and Significance* (Archaeopress, 2007).
10. McBrearty, S. & Tryon, C. A. in *Transitions Before the Transition: Evolution and Stability in the Middle Palaeolithic and Middle Stone Age* (eds Hovers, E. & Kuhn, S. L.) 257–277 (Springer, 2006).
11. McBrearty, S. Patterns of technological change at the origin of *Homo sapiens*. *Before Farming* **3**, 1–5 (2003).
12. Armitage, S. J. *et al.* The southern route ‘out of Africa’: evidence for an early expansion of modern humans into Arabia. *Science* **331**, 453–456 (2011).
13. Adler, D. S. *et al.* Early Levallois technology and the Lower to Middle Paleolithic transition in the Southern Caucasus. *Science* **345**, 1609–1613 (2014).
14. Clark, J. D. *et al.* Stratigraphic, chronological and behavioural contexts of Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* **423**, 747–752 (2003).
15. White, M., Scott, B. & Ashton, N. The Early Middle Palaeolithic in Britain: archaeology, settlement history and human behaviour. *J. Quat. Sci.* **21**, 525–541 (2006).
16. Tryon, C. A., McBrearty, S. & Texier, P.-J. Levallois lithics from the Kapthurin Formation, Kenya: Acheulian origin and Middle Stone Age diversity. *Afr. Archaeol. Rev.* **22**, 199–229 (2005).
17. DeBono, H. & Goren-Inbar, N. Note on a link between Acheulian handaxes and the Levallois method. *J. Israel Prehist. Soc.* **31**, 9–23 (2001).
18. Pappu, S. A Re-examination of the Palaeolithic Archaeological Record of Northern Tamil Nadu, South India (BAR International Series, 2001).
19. Haslam, M. *et al.* Late Acheulean hominins at the Marine Isotope Stage 6/5e transition in north-central India. *Quat. Res.* **75**, 670–682 (2011).
20. Mishra, S., Chauhan, N. & Singhvi, A. K. Continuity of microblade technology in the Indian Subcontinent since 45 ka: implications for the dispersal of modern humans. *PLoS ONE* **8**, e69280 (2013).
21. Blinkhorn, J. & Petraglia, M. D. in *Southern Asia, Australia and the Search for Human Origins* (eds Dennell, R. & Porr, M.) 64–75 (Cambridge Univ. Press, 2014).
22. Petraglia, M. *et al.* Middle Paleolithic assemblages from the Indian subcontinent before and after the Toba super-eruption. *Science* **317**, 114–116 (2007).
23. Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc. Natl Acad. Sci. USA* **110**, 10699–10704 (2013).
24. Pappu, S. *et al.* Early Pleistocene presence of Acheulian hominins in South India. *Science* **331**, 1596–1599 (2011).
25. Warrior, A. K. *et al.* A rock magnetic record of Pleistocene rainfall variations at the Palaeolithic site of Attirampakkam, southeastern India. *J. Archaeol. Sci.* **38**, 3681–3693 (2011).
26. Pappu, S., Akhilesh, K., Ravindranath, S. & Raj, U. Applications of satellite remote sensing for research and heritage management in Indian prehistory. *J. Archaeol. Sci.* **37**, 2316–2331 (2010).
27. Williams, M. A. J. *et al.* Environmental impact of the 73 ka Toba super-eruption in South Asia. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **284**, 295–314 (2009).
28. Raynaud, D. *et al.* Palaeoclimatology: the record for marine isotopic stage 11. *Nature* **436**, 39–40 (2005).
29. Schlebusch, C. M. *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
30. Berger, L. R., Hawks, J., Dirks, P. H., Elliott, M. & Roberts, E. M. *Homo naledi* and Pleistocene hominin evolution in subequatorial Africa. *eLife* **6**, e24234 (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements S.P. and K.A. thank the Sharma Centre for Heritage Education, the L. S. B. Leakey Foundation, the Earthwatch Institute, the Homi Bhabha Fellowships Council (S.P.: 2000–2002; K.A.: 2014–2016) and the ISRO-GBP program for funding various aspects of the research project, and the Archaeological Survey of India and Department of Archaeology, Government of Tamil Nadu, for issuing licenses. Y.G. benefited from an Institut Universitaire de France grant for field and analytical work. A.K.S. acknowledges the Department of Science and Technology and the Department of Atomic Energy, India, for a J. C. Bose national fellowship and for Raja Ramanna fellowships, respectively. H.M.R. was supported by the contingency grant of the J. C. Bose fellowship awarded to A.K.S. S.P. and K.A. thank M. Taieb for his encouragement.

Author Contributions K.A. and S.P. direct the project, are researching ATM and neighbouring sites and analysed the lithic artefacts; H.M.R., A.D.S. and A.K.S. were responsible for the luminescence sampling and dating; Y.G. analysed the geomorphology and palaeoenvironmental evidence at the site. All authors contributed to the writing of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.P. (pappu.shanti@gmail.com).

Reviewer Information *Nature* thanks K. Fitzsimmons, M. Petraglia and E. Rhodes for their contribution to the peer review of this work.

METHODS

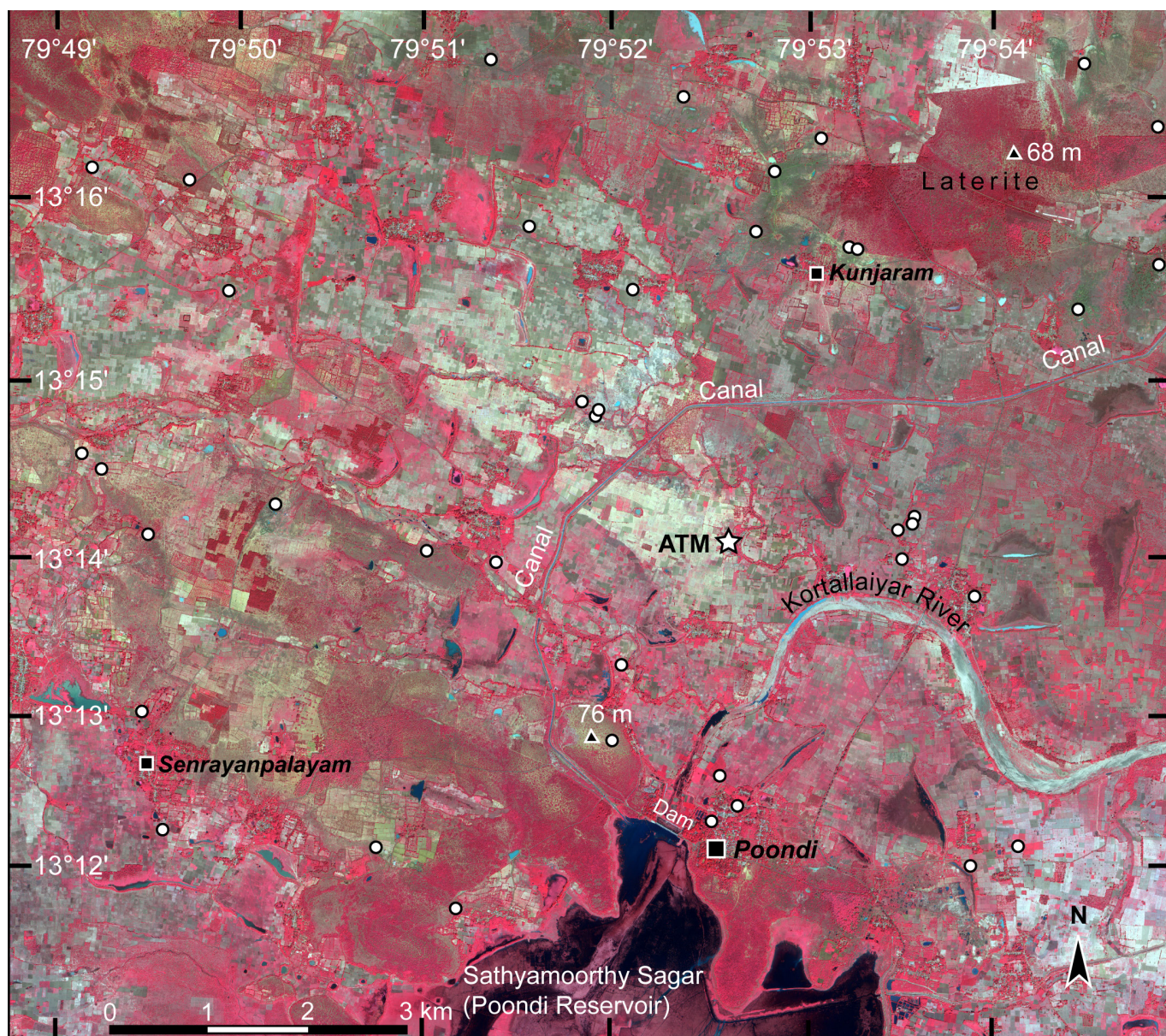
Optically stimulated luminescence (OSL) provides the burial age of archaeological sediments on the premise that at the time of deposition there was a minimal OSL signal in the constituent mineral grains. The reduction of OSL before burial occurs as a consequence of sediment grains being exposed to daylight during their transport and/or by human activity. On burial, exposure to daylight ceases and the re-accumulation of a luminescence signal in the mineral grains begins as a result of irradiation from ambient radioactivity. This continues until subsequent re-exposure or excavation for sampling. The measurement of total luminescence and its conversion to radiation dose units is possible using standard protocols (see Supplementary Information and references therein), and the corresponding radiation dose is called the palaeo-dose. When divided by the annual radiation dose, this gives the age (that is, time spent in the dark since the last exposure to daylight). The annual radiation dose is computed by the measurement of radioactivity concentrations produced in the burial environment by uranium, thorium, potassium and cosmic rays.

The luminescence age of a sedimentary deposit provides the time at which any archaeological artefacts contained within it were discarded and buried. In order

to date the sedimentary deposits containing Middle Palaeolithic assemblages at ATM, pIR-IRSL was preferred from among the range of existing luminescence methods, because of the anticipated antiquity of the stratigraphy and because the use of quartz OSL was hampered by the saturation of its luminescence signal (see Supplementary Information). Given the high solar irradiance, low sediment accumulation rate and indications from independent measurements on modern samples in similar environments in India (see Supplementary Information and references therein), it was reasonable to assume that the pIR-IRSL signal was bleached to low residual levels of less than a few Grays. This assumption is supported by the tight distribution of subsample ages in each sample (see Supplementary Information).

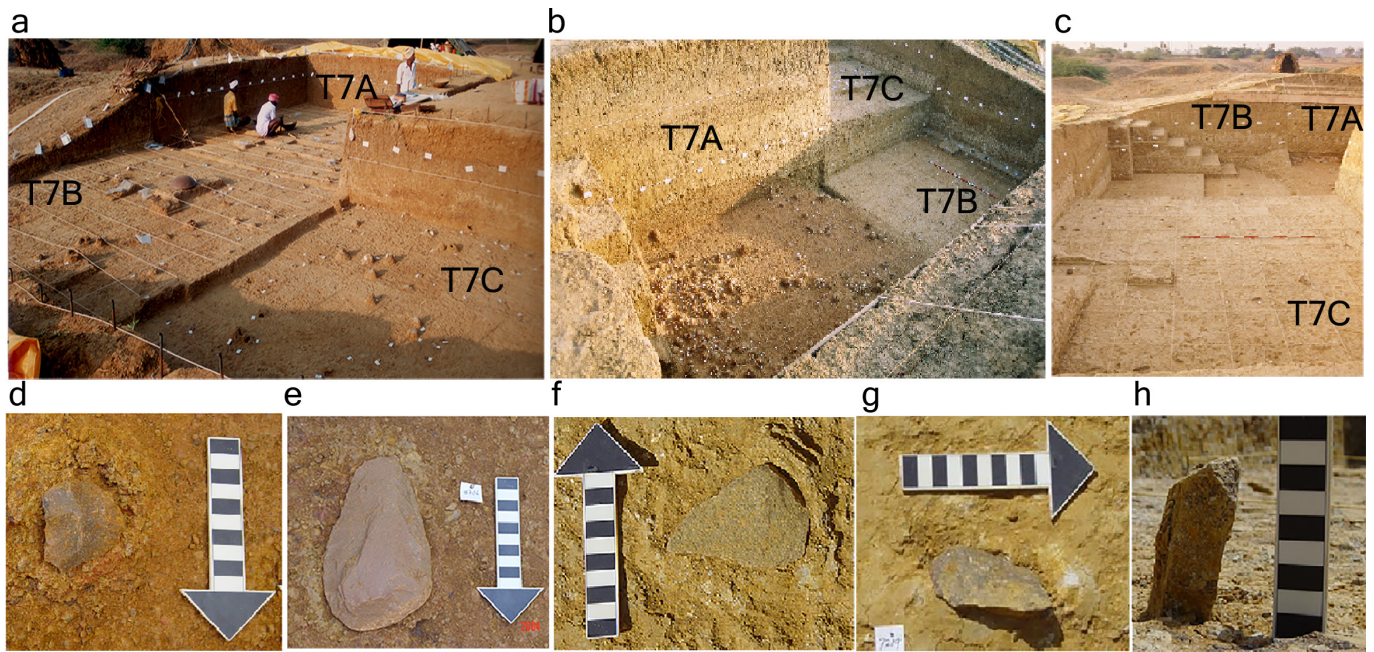
Data availability. All relevant data are included in the article and its Supplementary Information.

31. Pappu, S., Gunnell, Y., Taieb, M. & Akhilesh, K. Preliminary report on excavations at the Palaeolithic site of Attirampakkam, Tamil Nadu (1999–2004). *Man and Environment* **XXXIX**, 1–17 (2004).



Extended Data Figure 1 | Distribution of Acheulian, Middle Palaeolithic and Late Palaeolithic sites currently under investigation in the immediate vicinity of ATM, north-eastern Tamil Nadu. White circles: known prehistoric sites. Note their presence only on interfluvies or near smaller streams, and never on the riverbank of the largest local river (Kortallaiyar). Black squares: selected villages and towns. Red tone: vegetation (intensity increases with moisture content). Grey-to-green: dry crops and bare soil. Laterite caprock occurs in the north and forms

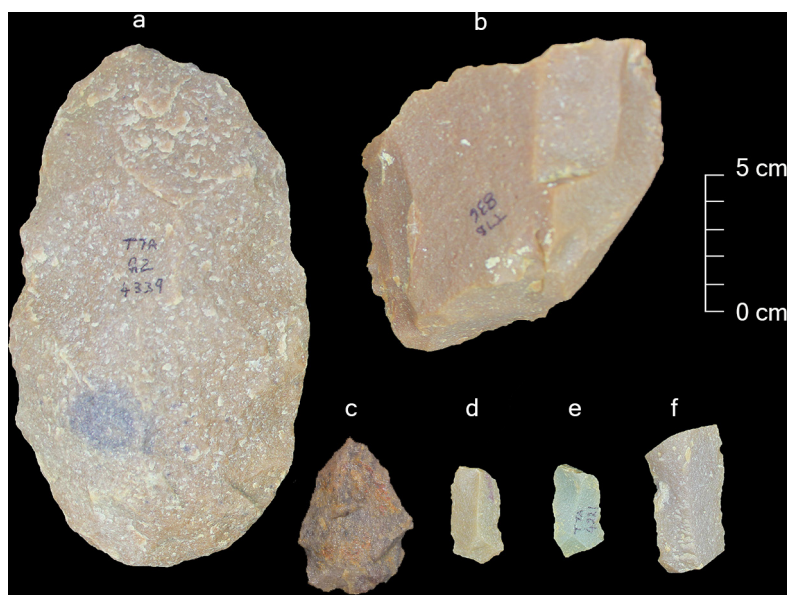
the high ground on the divide between the drainages of the Kortallaiyar and Arani rivers. Note that topographic gradients in the landscape can be inferred from the strike of the runoff-harvesting dams: some village reservoirs contain water, whereas others—which form dark crescentic patches in the image—are mostly dry. The land slopes perpendicular to dam orientation. False-colour IKONOS image, 1 m ground resolution. Reprinted from ref. 26, with permission from Elsevier.



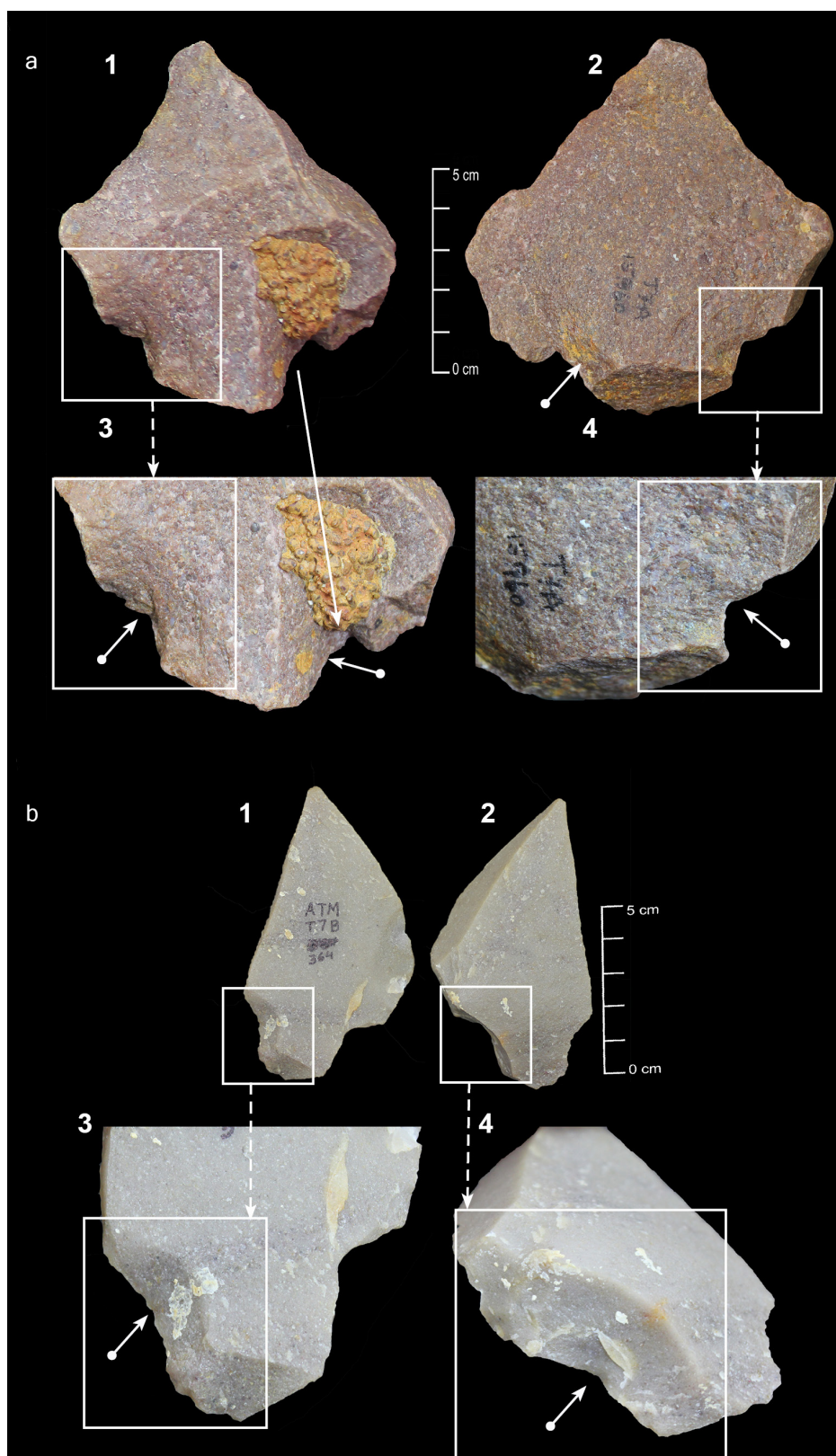
Extended Data Figure 2 | General view of excavated trenches and *in situ* artefacts. a–c, Range of views of trenches T7A, T7B and T7C. d, e, Artefacts from layer 5. f, g, Artefacts from layers 3 and 4. h, Artefact from layer 2. Arrow indicates north. Reprinted from ref. 31, with permission (a, c, e).



Extended Data Figure 3 | Artefacts from layer 5, trench T7A. a, b, Handaxes. c, Cleaver. d, Large flake tool. e, Blade core (see details in Extended Data Fig. 6b). f, g, Levallois cores. h, Levallois flake. i, Biface with a preferential flake removal. Arrow with solid circle indicates direction of flake scar removal.



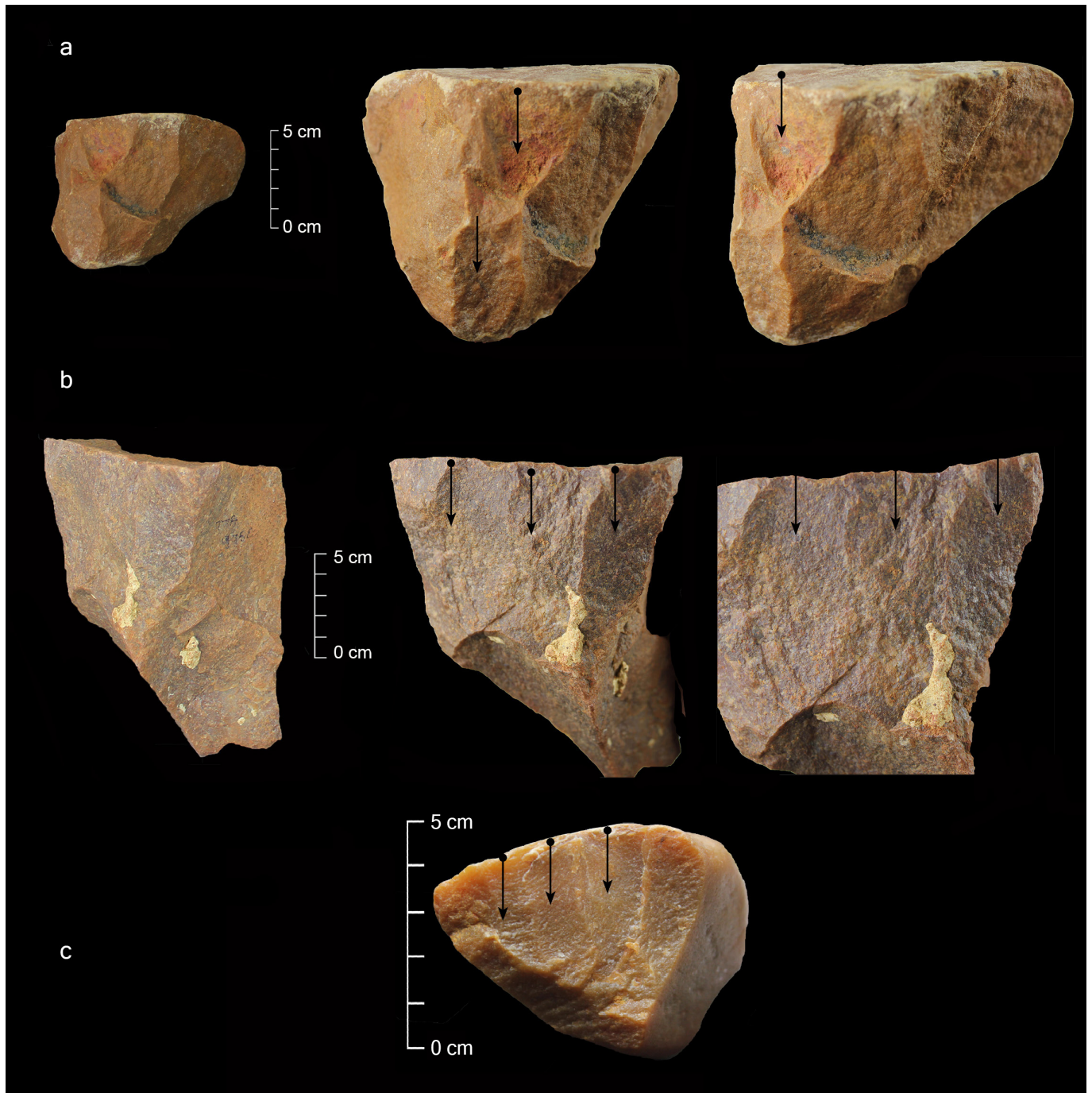
Extended Data Figure 4 | Representative artefacts from layers 2–4 and unit 5a, in trenches T7A, T7B and T7C. a, Biface, unit 5a. b, Blade core, layer 2. c, Retouched Levallois point, layer 4. d, e, Blades, layer 3. f, Blade, layer 2.



Extended Data Figure 5 | Close-up images of tanged artefacts.

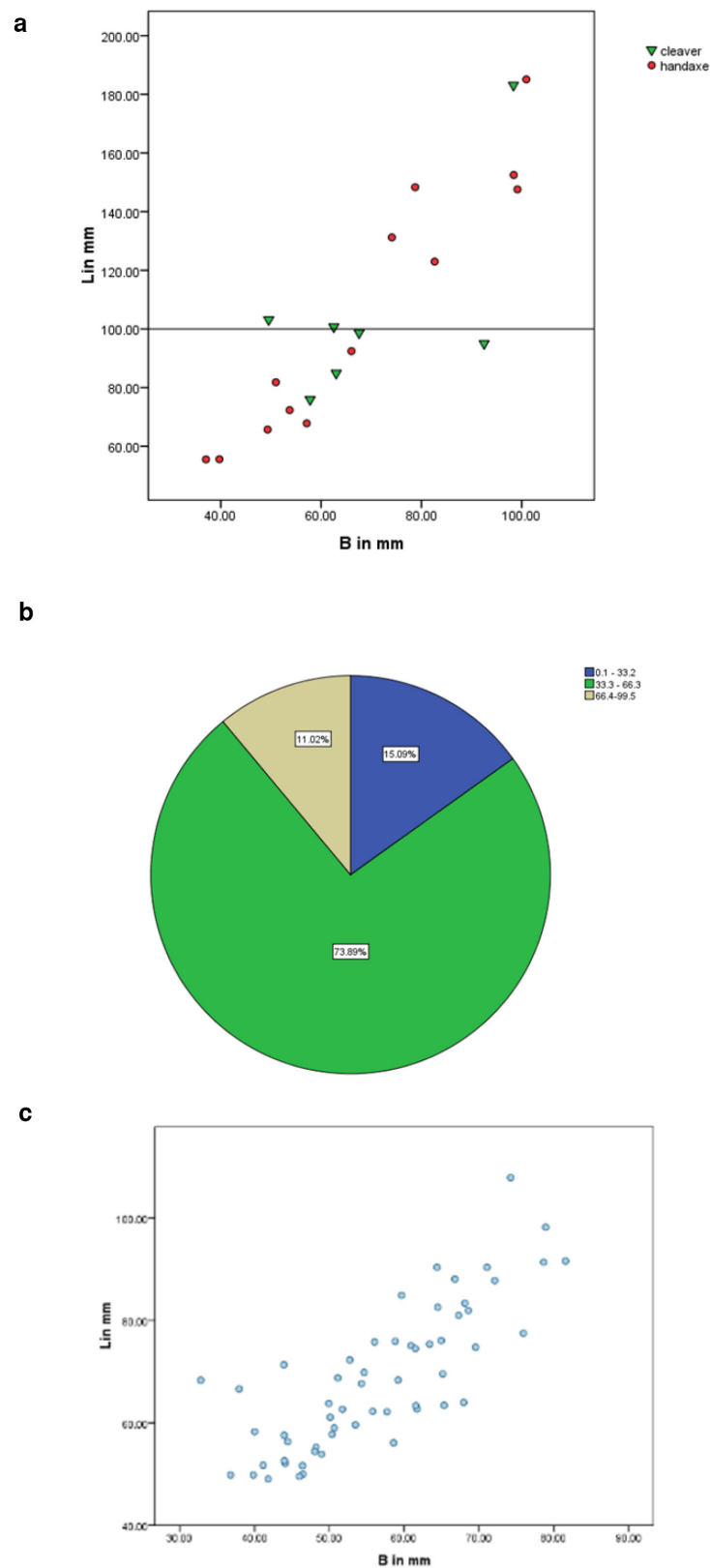
a, Tanged point from layer 5 (see Fig. 3j). **b**, Tanged point from layer 2 (see Fig. 4e). In **a** and **b**, 1 and 2 represent the dorsal and ventral faces of the artefact, respectively; 3 and 4 are close-up images of the dorsal and

ventral faces, respectively, showing details of the retouched areas. Boxes indicate enlarged areas of the tool depicting retouch. Arrows with white circle highlight retouch scars with a visible point of percussion.



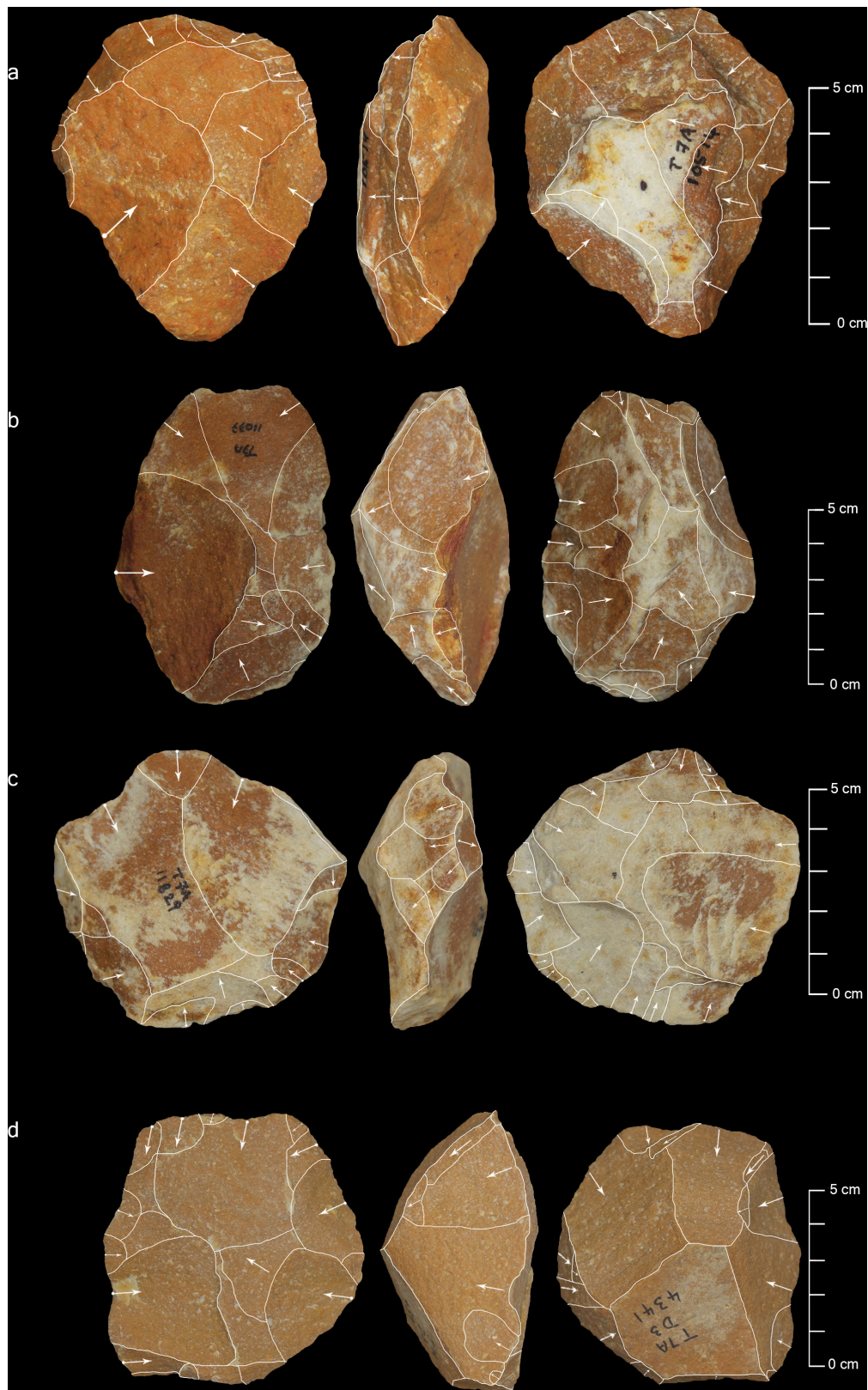
Extended Data Figure 6 | Blade cores with close-up images that illustrate generations of blade scar removals. a, Blade core from layer 5 (see Fig. 3a) (left); illustration of blade scar removals (right). **b,** Blade

core from layer 5 (see Extended Data Fig. 3e) (left); illustration of blade scar removals (right). **c,** Blade core on a cobble from layer 5. Black arrows indicate direction of percussion.



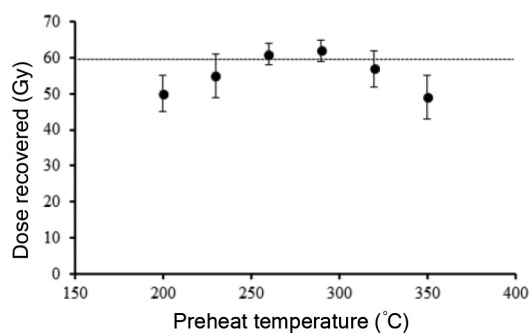
Extended Data Figure 7 | Details of artefact dimensions. **a**, Dimensions of bifaces from trench T7A. The horizontal line indicates the 100 mm mark, below which tools are considered diminutive⁹. **b**, Dimensions of small-flake tool component, showing binned values of length, from trench

T7A. Colours and groups indicate length range values in mm. Large cutting tools are not included owing to their negligible numbers. **c**, Dimensions of Levallois cores from trench T7A. L, length; B, breadth.



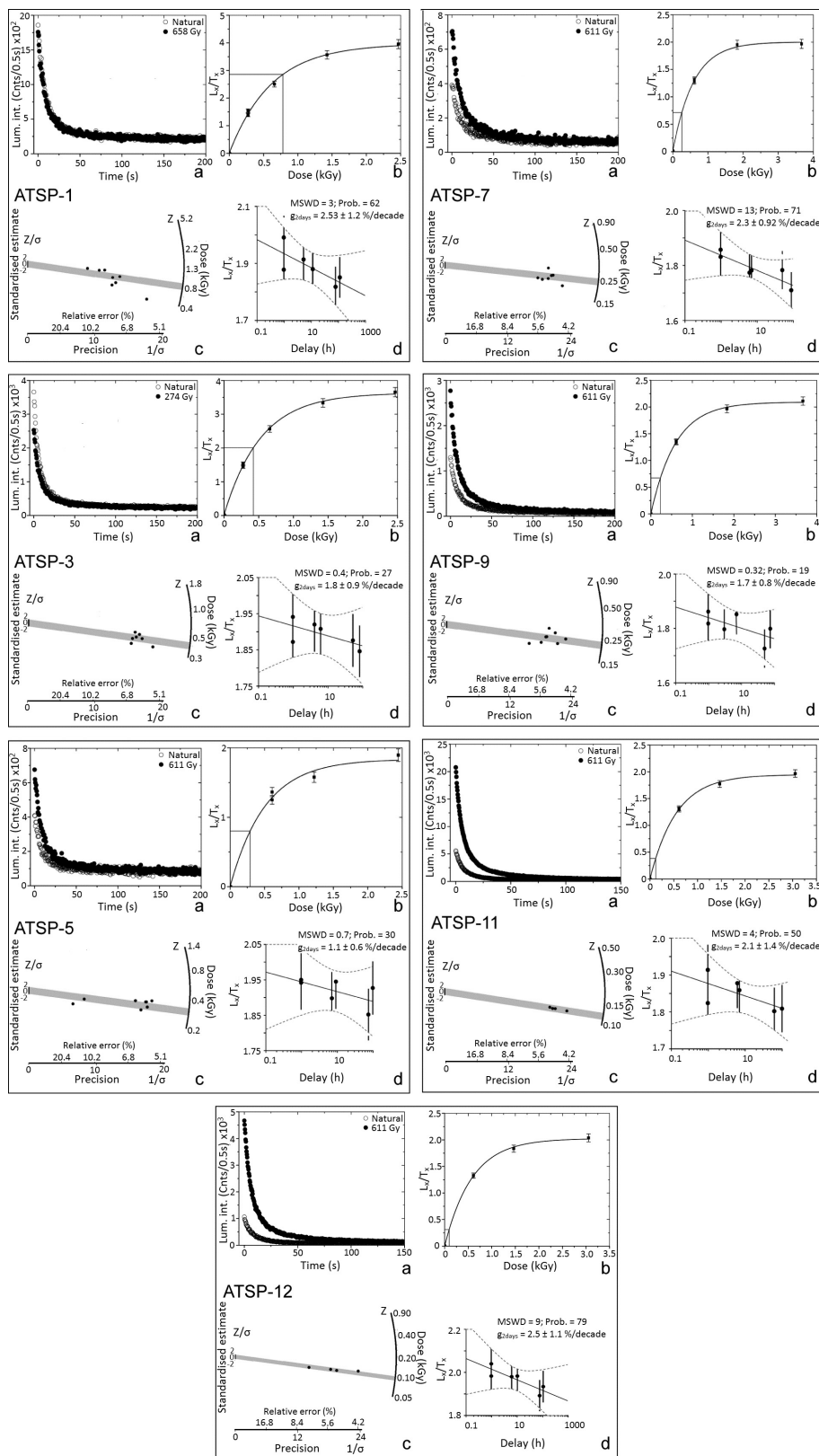
Extended Data Figure 8 | Levallois cores from ATM. **a**, Preferential Levallois flake core from trench T7A (layer 5, phase I). **b**, Preferential Levallois flake core from trench T7A (layer 5, phase I). **c**, Recurrent unidirectional Levallois core from trench T7A (layer 5, phase I).

d, Recurrent centripetal Levallois core from trench T7A (layer 3, phase II). Superimposed white lines indicate flake scars. Arrows indicate direction of flake removals; arrows with solid circle indicate the point of percussion.



Extended Data Figure 9 | Pre-heat plateau test for sample

ATSP#2013/12L1. A beta dose of 60 Gy was given after 8 h of bleaching of a fresh natural sample under a solar stimulator. Each data point shown represents the arithmetic mean of three aliquots for each pre-heat step. The recovered doses for 260 °C, 290 °C and 320 °C fell within the 5% of the total variation, that is within 57–63 Gy for an administered dose of 60 Gy. Error bars are the standard error of the mean of the three aliquots for each pre-heat temperature.



Extended Data Figure 10 | See next page for caption.

Extended Data Figure 10 | Results of pIR-IRSL analyses at ATM. The panels show **a–d** separately for ATSP-1, ATSP-3, ATSP-5, ATSP-7, ATSP-9, ATSP-11 and ATSP-12. **a**, Typical feldspar shine-down curve. Open circles, natural irradiation; black circles, after beta irradiation. **b**, Typical growth curve. **c**, Radial plot representing scatter in estimated palaeo-doses. **d**, Typical g -value data. In **c**, the x axis represents precision ($1/\sigma$) and the y axis represents the standardized estimate (Z/σ) with respect to the mean value (here, Z_0); the circular scale represents estimate values of Z (or dose) in Gy. Points with a larger x value have higher precision. The grey band shows the 2σ envelope of the calculated mean. In **d**, the probability value is the probability of the χ^2 distribution and the dashed curves represents the 2σ hyperbolic error envelope, defining the confidence region around

the best-fit line. On average, the probability value should be equal to 50%. Larger values indicate that the overall scatter of the data (around the mean regression line) is smaller than the uncertainties of each data point. A smaller probability demonstrates that the scatter is larger than the uncertainties of each data point, and the mean square weighted deviation (MSWD) measures the goodness of fit of the model to the data. $MSWD = 1$ when the data fit a univariate normal distribution as a function of t or $\log_{10}(t)$. $MSWD > 1$ when the observed scatter exceeds predicted analytical uncertainty (data are over-dispersed). $MSWD < 1$ when the observed scatter is less than the predicted analytical uncertainty (data are under-dispersed).

The honeycomb maze provides a novel test to study hippocampal-dependent spatial navigation

Ruth A. Wood^{1*}, Marius Bauza^{1*}, Julija Krupic^{2,3*}, Stephen Burton¹, Andrea Delekate⁴, Dennis Chan⁵ & John O'Keefe^{1,3}

Here we describe the honeycomb maze, a behavioural paradigm for the study of spatial navigation in rats. The maze consists of 37 platforms that can be raised or lowered independently. Place navigation requires an animal to go to a goal platform from any of several start platforms via a series of sequential choices. For each, the animal is confined to a raised platform and allowed to choose between two of the six adjacent platforms, the correct one being the platform with the smallest angle to the goal-heading direction. Rats learn rapidly and their choices are influenced by three factors: the angle between the two choice platforms, the distance from the goal, and the angle between the correct platform and the direction of the goal. Rats with hippocampal damage are impaired in learning and their performance is affected by all three factors. The honeycomb maze represents a marked improvement over current spatial navigation tests, such as the Morris water maze^{1–3}, because it controls the choices of the animal at each point in the maze, provides the ability to assess knowledge of the goal direction from any location, enables the identification of factors influencing task performance and provides the possibility for concomitant single-cell recording.

The hippocampal formation generates a cognitive map of a familiar environment that supports the ability of an animal to identify its location, respond to changes in the environment and navigate to desirable locations or avoid undesirable ones⁴. These functions are supported by cells coding for location (place cells)⁵, heading direction (head direction cells)^{6,7}, distance in a particular direction (grid cells)⁸ and distance from the boundaries of an environment (boundary cells)^{9,10}; these functions have previously been reviewed^{11,12}.

Several tasks are routinely used to test spatial navigational learning. These include the T maze and Y maze¹³, Olton radial arm maze^{14,15}, Barnes maze¹⁶ and Morris water maze^{1–3}. All have disadvantages as tests of spatial navigation and memory, and for concomitant recording of spatial cell activity. The first four do not force animals to use a single identifiable spatial navigational strategy, because they can all be learned using directional- or object-heading strategies as well as place learning. The Morris water maze overcomes this indeterminacy because it requires the animal to approach a hidden goal from a variety of directions; the animal must therefore head in different directions and approach different cues on each individual trial. However, the unlimited number of choices at each location, the lack of independence between

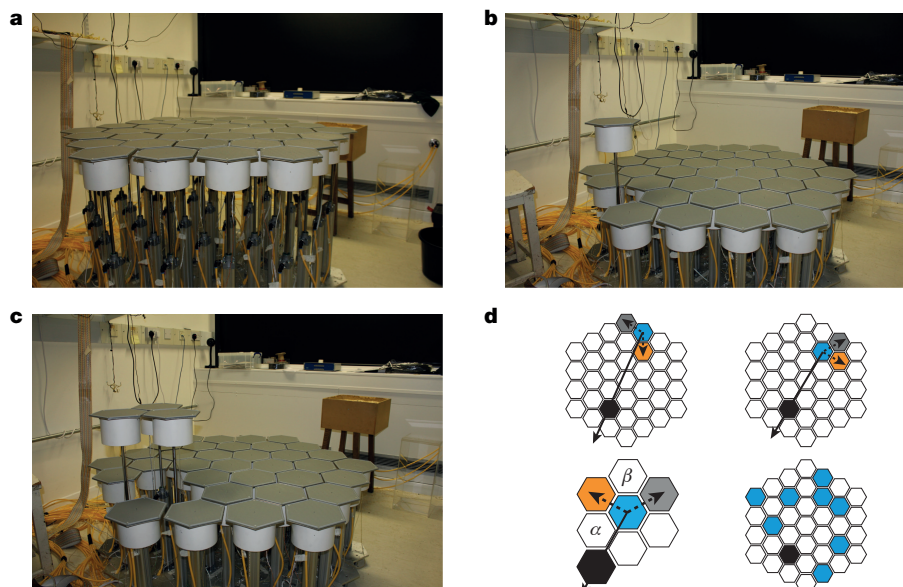


Figure 1 | The honeycomb maze. a–c, Honeycomb maze with all (a), one (b) and three (c) platforms raised. d, Schematic navigation paradigm. Top left, at any given starting location (blue) two choices are offered: the correct choice (orange) possesses a smaller angle with respect to the heading direction towards the goal (black) than does the incorrect choice (grey). Top right, the previously chosen platform becomes the

new ‘occupied’ platform (blue), and two further platforms are presented as choices. Bottom left, each choice is described by two angles: angle α between the correct choice and goal-heading direction, and angle β between the correct and incorrect choices. Bottom right, illustration of eight potential starting platforms (blue).

¹Sainsbury Wellcome Centre, UCL, London W1T 4JG, UK. ²Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge CB2 3DY, UK. ³Cell & Developmental Biology, UCL, London WC1E 6BT, UK. ⁴DZNE German Centre for Neurodegenerative Diseases, Bonn 53127, Germany. ⁵Department of Clinical Neurosciences, University of Cambridge, Cambridge CB2 2PY, UK.

*These authors contributed equally to this work.

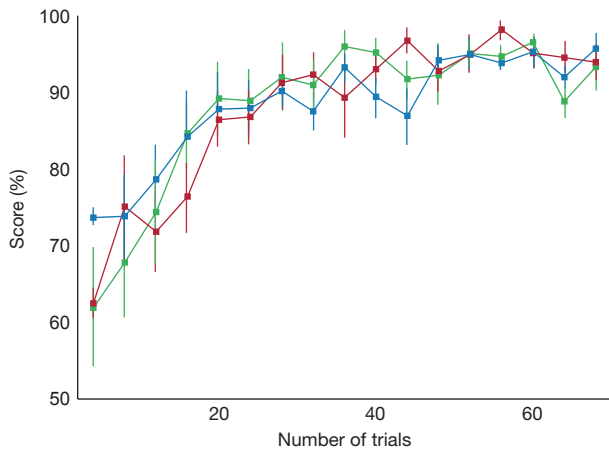


Figure 2 | Learning to navigate the honeycomb maze. Performance of three groups of rats; data are mean \pm s.e.m. and each data point is representative of four trials. $n = 9$ unoperated controls (blue), 8 controls with sham hippocampal lesions (red) and 8 controls with sham medial entorhinal lesions (green). All rats completed between 68 and 72 trials, over 13–17 days; data are shown for the first 68 trials. Unoperated controls did six trials per day, sham-operated controls did four trials per day. Performances were similar for all groups ($F_{2,22} < 0.001$, $P > 0.999$, two-way mixed ANOVA).

successive choices and the inhomogeneity of behaviour across the environment once the animal has learned the task (see Supplementary Discussion 1, 2) all present problems for studies that use the Morris

water maze; the honeycomb maze described here has been designed to overcome these difficulties.

Here we describe the honeycomb maze, the performance of control rats during the navigational task, the factors that affect their performance as well as the performance of rats with hippocampal lesions, and, as proof of principle, an example of a place cell recorded on the maze is included in Extended Data Fig. 1.

The honeycomb maze consists of 37 tessellated hexagonal platforms each fixed atop a pneumatic tube, which enables it to be raised independently of the others. Figure 1 shows the maze in a variety of configurations: all platforms raised (Fig. 1a), a single platform raised (Fig. 1b) or three adjacent platforms raised (Fig. 1c). The objective is to reach a specific goal platform from eight or nine starting locations (Fig. 1d, bottom right, Extended Data Fig. 2) by making a series of binary choices between two platforms adjacent to the currently occupied platform (Fig. 1d, top left and top right, Supplementary Video 1). The ‘correct’ choice is the platform with the smallest angle to the goal direction—in vector terminology, the one with the smallest dot or inner product with respect to the goal-direction vector. The sequence continues until the goal platform is reached, where after a short delay the animal is rewarded with a single Cheerio.

First, three groups of male Lister hooded rats were tested on a spatial navigation task using the honeycomb maze (Methods): a control group that had not been operated on and two control groups that had been operated on; one group received sham surgical procedures in the medial entorhinal cortex and the other sham hippocampal lesions. These three groups were used as controls for a further experiment with rats that had ibotenic-acid-induced lesions in the hippocampus, described below. All three control groups rapidly learned the spatial navigation task, and

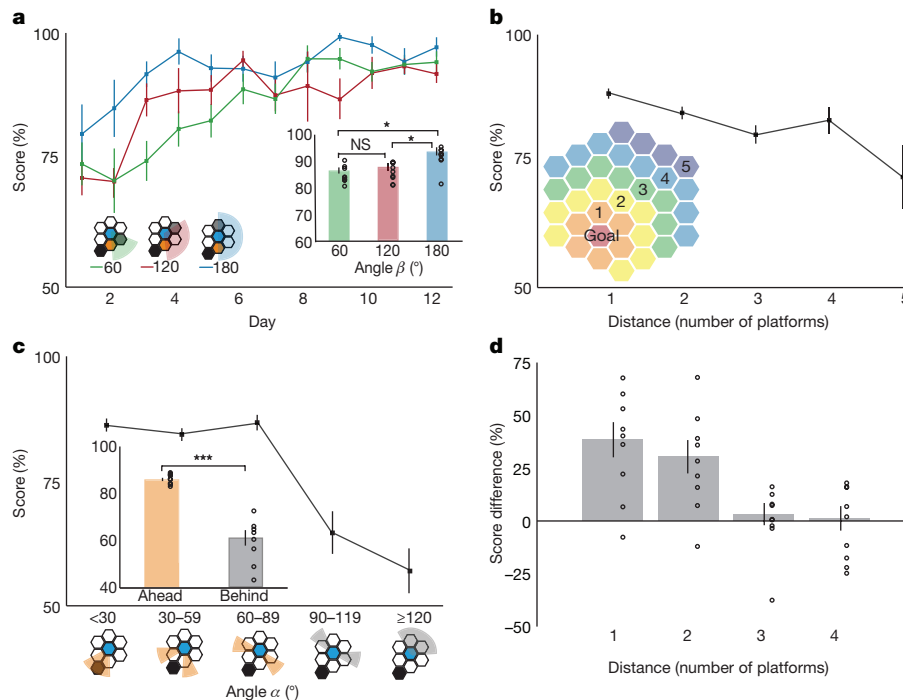


Figure 3 | Factors that affect maze performance in controls. Three factors influence performance in unoperated control rats ($n = 9$ rats).

a, Performance improved with increasing platform separation (performance versus angle β : $F_{2,16} = 8.850$, $P = 0.003$) and with experience (performance versus testing day: $F_{11,88} = 11.361$, $P < 0.001$), but there was no significant interaction between angle β and experience ($F_{22,176} = 1.438$, $P = 0.102$). All two-way repeated measures ANOVA. Inset, mean performance for different values of angle β across all days, with pairwise comparisons: 60° versus 180°, mean difference = 7.0%, $P = 0.034$; 120° versus 180°, mean difference = 5.9%, $P = 0.044$; 60° versus 120°, mean difference = 1.1%, $P = 1.000$; post hoc Bonferroni test. Six trials per day. **b**, Performance decreased with increasing distance of choice from goal,

measured as in the inset ($F_{3,24} = 3.707$, $P = 0.025$, one-way repeated measures ANOVA). **c**, Performance decreased with increasing angle to goal (angle α , $F_{4,32} = 20.670$, $P < 0.001$, one-way repeated measures ANOVA). Inset, mean scores were higher for choices that were ‘ahead’ (angle $\alpha < 90^\circ$, orange) than they were for choices that were ‘behind’ (angle $\alpha > 90^\circ$, grey) ($t_8 = 6.620$, $P < 0.001$, two-sided paired t -test). **d**, There was an interaction between distance from goal and angle to goal (angle α) for adjacent platform choices ($F_{3,24} = 9.133$, $P < 0.001$, two-way repeated measures ANOVA). The y axis shows the difference in mean performance for ‘ahead’ and ‘behind’ choices. For **a–d**, error bars indicate s.e.m., * $P < 0.05$, *** $P < 0.001$.

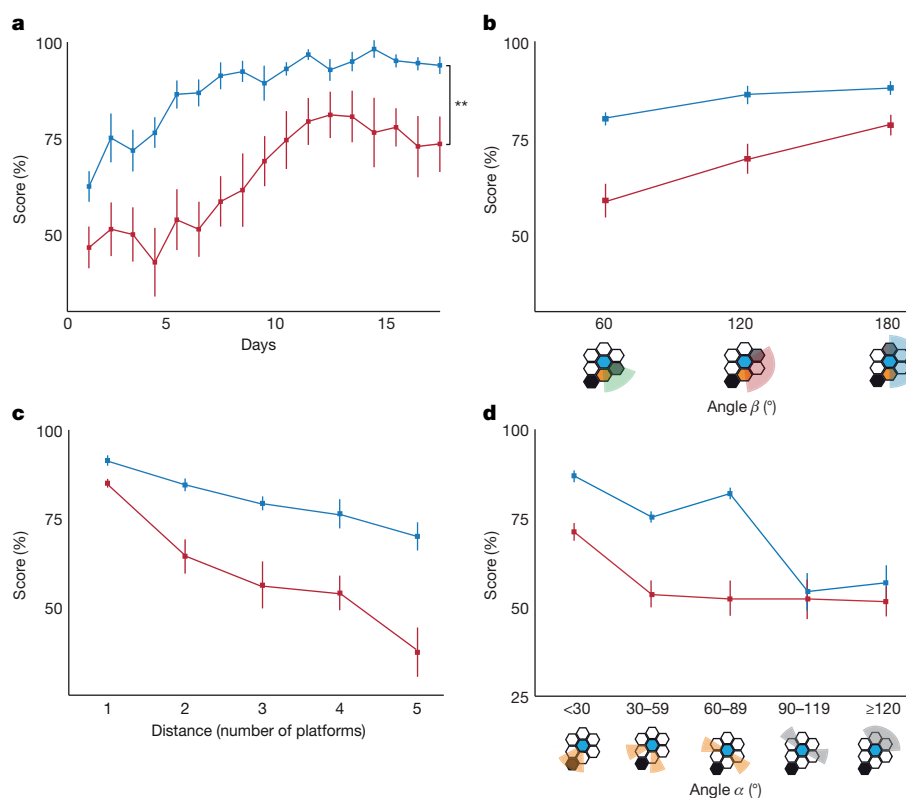


Figure 4 | Performance of rats with hippocampal lesions on the honeycomb maze. **a**, Rats with hippocampal damage (red, $n = 8$) are significantly impaired when learning the task compared to controls with sham hippocampal lesions (blue, $n = 8$) ($F_{1,14} = 10.240$, $P = 0.006$, two-way mixed ANOVA), four trials per day. **b–d**, Rats with hippocampal damage (red, $n = 8$) are significantly more influenced than controls (blue, $n = 8$) by

separation between choice platforms (**b**, angle β ; $F_{2,28} = 40.024$, $P < 0.001$, two-way mixed ANOVA), distance from the goal (**c**, $F_{4,56} = 34.740$, $P < 0.001$, two-way mixed ANOVA) and angle to goal (**d**, angle α ; $F_{2,3,32,8} = 28.812$, $P < 0.001$, two-way mixed ANOVA). For **a–d**, error bars indicate s.e.m., ** $P < 0.01$.

after 28 trials all rats achieved a mean score of greater than 90% correct choices over four consecutive trials. There was no difference in performance between the three groups, which suggests that performance on the maze is consistent and reliable (Fig. 2).

Several parameters determined performance in the unoperated control rats: (i) the angular difference between the direction to the correct platform and the goal (angle α); (ii) the smallest angular separation between the two choice platforms, independent of the direction of the goal (angle β); and (iii) the distance of the choice from the goal (Supplementary Discussion 3). The testing schedule minimized the interaction between these variables so that the effect of each could be investigated (Methods, Extended Data Fig. 2). The rats performed best when choice platforms were separated by two others (angle $\beta = 180^\circ$; $92.8 \pm 1.1\%$ (mean \pm s.e.m.) correct choices overall) and deteriorated when only one (angle $\beta = 120^\circ$) or no (angle $\beta = 60^\circ$) platform separated them ($86.9 \pm 1.2\%$ and $85.8 \pm 1.4\%$ respectively). There was a significant effect of both angle β and day of testing on performance, but no interaction between these two variables (Fig. 3a). The differences between two-platform separations (angle $\beta = 180^\circ$) and the rest (angle $\beta = 120^\circ$ or 60°) were significant (Fig. 3a, inset).

Performance decreased with increasing distance from the goal (Fig. 3b, Supplementary Discussion 3); the percentage of correct choices ranged from $88.4 \pm 1.0\%$, when platforms were adjacent to the goal, to $71.8 \pm 6.1\%$, when platforms were five platforms away from the goal. Performance also decreased as an inverse function of the angular separation between the direction to the goal and the correct platform (angle α); the percentage of correct choices ranged from $86.2 \pm 1.3\%$ for angles of 0° – 29° , to $61.1 \pm 3.1\%$ for angles greater than or equal to 90° (Fig. 3c). Rats performed significantly better when the direction of the correct platform was less than 90° from that of the goal (Fig. 3c, inset). Notably, even when the correct choice was 90° or greater from the goal direction,

all rats performed with rates of success above those that would be achieved by chance ($t_8 = 3.156$, $P = 0.013$, one sample t -test).

A multiple regression analysis indicated that angle α , angle β and distance were all significant predictors of performance (Extended Data Table 1). Between them, these variables predicted 5.5% of the variance ($R^2 = 0.055$, $F_{3,391} = 7.608$, $P < 0.001$); a large proportion of the remaining variance in performance is accounted for by experience (Fig. 2).

Finally, there was evidence of an interaction between the effects of angle α and the distance from the goal, at least when choice platforms were adjacent to one another. The angle-to-goal effect decreases as a function of the distance from the goal, such that performance improved with larger angles at greater distances (Fig. 3d).

In a further experiment, learning on the honeycomb maze was compared between rats with ibotenic-acid-induced lesions of the hippocampus and sham-operated control rats. Hippocampal damage ranged from 48% to 94%; preserved tissue was observed primarily in the ventral hippocampus and small amounts of incidental damage to the caudate nucleus and putamen were found in all rats. Minor additional damage to the dorsal subiculum, medial geniculate nucleus and pre- and parasubiculum was found in a subset of rats (Methods, Extended Data Fig. 3).

Rats with hippocampal lesions were significantly deficient in learning the task, relative to controls (Fig. 4a). As in the unoperated control group, performance in rats with hippocampal lesions and in the sham-operated controls was related to three variables (Fig. 4b–d): separation between choice platforms (angle β), distance from the goal and angle of the correct choices to the goal direction (angle α). There was a significant interaction between all three variables and lesion status (angle $\beta \times$ lesion status, $F_{2,28} = 6.981$, $P = 0.003$; distance \times lesion status, $F_{4,56} = 4.999$, $P = 0.002$; angle $\alpha \times$ lesion status, $F_{2,3,32,8} = 8.431$,

$P = 0.001$; two-way mixed ANOVAs). Behavioural performance decreased as hippocampal damage increased, although this was not significant, possibly owing to a floor effect as half of the rats with lesions scored at chance level (Extended Data Fig. 4). Rats with hippocampal lesions were also slower to make choices ($F_{1,14} = 11.103$, $P = 0.005$; Extended Data Fig. 5). Latencies were significantly longer for incorrect choices than correct choices across all rats ($F_{1,14} = 23.839$, $P < 0.001$), which suggests that longer latencies reflect the uncertainty of the rats; this effect was larger for rats with lesions ($F_{1,14} = 4.956$, $P = 0.043$, two-way mixed ANOVA).

In summary, the performance of rats across all three control groups was comparable (Fig. 2) even when the rats were tested by different experimenters on different occasions, which shows the consistency and reproducibility of the task. Performance was affected by three variables. Success was reduced with increased distance of the choice to the goal and by increased deviation of the best-choice platform from the direction vector to the goal (angle α), and was improved by an increase in the angle between the two choice platforms (angle β). Both direction and distance factors were noted in early research on maze learning (Supplementary Discussion 3), and it has been proposed that the direction factor is generated by the hippocampal cognitive map⁴. The ability of control rats to identify the better of two directions even when neither is aligned with the goal suggests that the brain is capable of vector computations. Rats with hippocampal damage performed significantly worse for all three factors, which suggests that these vectors are computed in the hippocampus itself. An alternative explanation is that the computation is performed in another brain region—such as the parietal cortex—that requires input from hippocampal place cells that encode the current location of the animal, the location of the goal and the choice platform locations. A simple vector schema can account for these data (Extended Data Fig. 6), as has previously been proposed^{17,18}.

Finally, the honeycomb maze makes an ideal environment for correlating the activity of place cells with spatial navigation performance (Extended Data Fig. 1), which is not easily accomplished in the water maze. The honeycomb maze therefore represents an improvement on the Morris water maze, which is the current gold standard for testing hippocampal-dependent spatial navigation (Supplementary Discussion 2).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 20 June; accepted 4 December 2017.

Published online 24 January 2018.

1. Morris, R. G. M. Spatial localization does not require the presence of local cues. *Learn. Motiv.* **12**, 239–260 (1981).
2. Morris, R. G., Garrud, P., Rawlins, J. N. & O'Keefe, J. Place navigation impaired in rats with hippocampal lesions. *Nature* **297**, 681–683 (1982).
3. Morris, R. Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Methods* **11**, 47–60 (1984).

4. O'Keefe, J. & Nadel, L. *The Hippocampus as a Cognitive Map* (Oxford Univ. Press, 1978).
5. O'Keefe, J. & Dostrovsky, J. The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Res.* **34**, 171–175 (1971).
6. Ranck, J. B. Head-direction cells in the deep cell layers of the dorsal presubiculum in freely moving rats. *Abstr. Soc. Neurosci.* **10**, 599 (1984).
7. Taube, J. S., Muller, R. U. & Ranck, J. B. Jr. Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis. *J. Neurosci.* **10**, 420–435 (1990).
8. Hafting, T., Fyhn, M., Molden, S., Moser, M. B. & Moser, E. I. Microstructure of a spatial map in the entorhinal cortex. *Nature* **436**, 801–806 (2005).
9. Lever, C., Burton, S., Jeewajee, A., O'Keefe, J. & Burgess, N. Boundary vector cells in the subiculum of the hippocampal formation. *J. Neurosci.* **29**, 9771–9777 (2009).
10. Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B. & Moser, E. I. Representation of geometric borders in the entorhinal cortex. *Science* **322**, 1865–1868 (2008).
11. O'Keefe, J. in *The Hippocampus Book* (eds Morris, R. et al.) Ch. 8, 471–544 (Oxford Univ. Press, 2007).
12. Moser, M. B., Rowland, D. C. & Moser, E. I. Place cells, grid cells, and memory. *Cold Spring Harb. Perspect. Biol.* **7**, a021808 (2015).
13. Dudchenko, P. A. How do animals actually solve the T maze? *Behav. Neurosci.* **115**, 850–860 (2001).
14. Olton, D. S., Walker, J. A. & Gage, F. H. Hippocampal connections and spatial discrimination. *Brain Res.* **139**, 295–308 (1978).
15. Olton, D. S. & Samuelson, R. J. Remembrance of places passed: spatial memory in rats. *J. Exp. Psychol. Anim. Behav. Process.* **2**, 97–116 (1976).
16. Barnes, C. A. Memory deficits associated with senescence: a neurophysiological and behavioral study in the rat. *J. Comp. Physiol. Psychol.* **93**, 74–104 (1979).
17. O'Keefe, J. in *Brain and Space* (ed. Paillard, J.) 273–295 (Oxford Univ. Press, 1991).
18. O'Keefe, J. in *Language and Space* (eds Bloom, P. et al.) 277–316 (MIT Press, 1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Bertelli, A. Hastings, D. Howett, N. Khan, P. Mumford, A. O'Leary, B. Potter, S. Richards and R. Wu for their contribution to this work, and D. Farquharson and D. Halpin for their input to the design, building and testing of maze prototypes. This work was supported by grants from the Wellcome Trust and the Gatsby Charitable Foundation to J.O. R.A.W. is an MRC Clinical Research Training Fellow, J.K. is a Wellcome Trust/Royal Society Sir Henry Dale Fellow and is supported by the Kavli Foundation Dream Team project and the Isaac Newton Trust. D.C. is funded by the Cambridge NIHR Biomedical Research Centre and by the Wellcome Trust.

Author Contributions J.O. conceived the maze and the study. J.O., M.B., J.K. and S.B. were instrumental in designing, building and testing prototypes of the maze. J.K. and M.B. designed the custom-made software used to operate the maze. R.A.W. designed testing schedule 1 for the control experiment, and A.D. and J.O. designed testing schedule 2 for the lesion experiment. A.D. and S.B. performed the hippocampal lesion and sham lesion surgeries. R.A.W. acquired the behavioural data. R.A.W. and A.D. performed the histology for the lesion experiment, and R.A.W. measured hippocampal lesion volumes. R.A.W. conducted the data management and performed the statistical analyses. J.K., S.B. and J.O. collected the single-unit data and J.K. and J.O. analysed these data. J.O. and R.A.W. wrote the manuscript, with contributions to later drafts from all other authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.O. (j.okeefe@ucl.ac.uk).

METHODS

Maze description. The honeycomb maze consists of 37 tessellated hexagonal platforms (11.5 cm each side) in an overall hexagonal configuration (total maze diameter 145.5 cm, five platforms across with 1-cm gaps between each platform). Each platform is fixed atop a pneumatic tube that enables it to be raised independently of the others. Each raised platform is 81.5 cm above the base (49 cm when lowered). Each platform consists of three layers: the movable top is made of 3-mm thick acrylonitrile butadiene styrene (ABS) with a pinseal finish, and is connected to the bottom layer by six microswitches that register the presence of an animal on the platform; the middle layer is made of aluminium, which is grounded to reduce electrical artefacts. Custom-made software that was written in Labview is used to monitor the platform switches, which enables the raising and lowering of platforms on the basis of the animal's location. Only unoccupied platforms are moved. Plastic skirts around the top of the pneumatic tubes protect the wiring and tubing under the platform. The maze sits within a standard laboratory, which gives the animal access to abundant extra-maze cues for location and direction (Fig. 1a–c).

Animals and surgical procedures. Four groups of rats weighing between 310 and 374 g were trained: three control groups and a group with hippocampal lesions. All rats were male Lister hooded rats purchased from Charles River Laboratories and aged between 12 and 16 weeks at the start of behavioural testing. Rats were all housed in open caging along with their littermates, in groups of three to four. All animal experiments were carried out in accordance with British Home Office Regulations (UK Animals Scientific Procedures Act 1986; Project License PPL 70/8202 to J.O.). Study protocols were in accordance with the terms of the Project License, which was reviewed by the Animal Welfare and Ethical Review Board at University College London.

The first group consisted of nine unoperated control rats. The second group comprised eight control rats with sham hippocampal lesions, and the third group comprised eight rats with sham lesions in the medial entorhinal cortex. Group sizes were determined on the basis of extensive prior experience with maze studies. In the operated controls, sham lesions that caused no neural damage were made by insertion of a borosilicate glass micropipette (World Precision Instruments) without injection into fourteen sites per hemisphere for the sham hippocampal lesions (Extended Data Table 2), and eight sites per hemisphere for the sham medial entorhinal cortex lesions (Extended Data Table 3). In the final group of eight rats, pressure injections of between 50–80 nl of ibotenic acid ($10 \mu\text{g}\mu\text{l}^{-1}$) were made via a glass micropipette at fourteen injection sites in each hemisphere to lesion the hippocampus (Extended Data Table 2). Surgeries were performed under sterile conditions under isoflurane anaesthesia and the rats were given analgesics (Metacam) for three days after the operation.

For the lesion study, rats were randomized into an intervention group, which received hippocampal lesions, or an operated control group, which received sham hippocampal lesions. Randomization was stratified to ensure that littermates were equally distributed between the two groups.

Behavioural training. After one week of recovery, all rats were food-restricted and reduced to 90% of their free-feeding body weight over a two-week period, before training. During this time, they were handled daily, and during the final three days they were placed on a holding platform in the testing room where they were habituated to the sounds made by the maze platforms being raised and lowered.

There was no period of maze exploration before training on the task. At the beginning of each trial, the rat was placed on one of eight or nine starting platforms and after a delay of four seconds, two of the six adjacent platforms were raised (Fig. 1d). When the rat chose one of these two platforms, the previously occupied and non-chosen platforms were lowered and after an interval of four seconds, two new platforms were raised. Neither of these was the platform from which the rat had just come (with one exception), which ruled out the strategy of avoiding that platform as a potential solution. If a choice was not made within one minute, the rat was gently guided onto a platform, which was alternatively chosen as correct or incorrect—ruling this out as a source of information about the correct choice. The choice was then scored as incorrect, and the next choice in the sequence initiated. If the rat had not reached the goal within five minutes, the trial was terminated and the rat guided to the goal platform and given a reward. On some 'choices', the rat was presented with only a single platform and was required to move to that platform. This forced choice occasionally became necessary to eliminate the strategy of avoiding a platform that had immediately previously been occupied.

Platform choices varied along several dimensions: (i) in the angle between the best choice platform and the heading direction to the goal, which varied from 0° to 135° (angle α , Fig. 1d), (ii) in the angle between the two platforms (angle β , Fig. 1d), with choices that ranged from neighbouring platforms to platforms separated by two others and (iii) in the distance of the platform from the goal in terms of the number of platforms (1–5) to be traversed on the direct route to the goal (Fig. 3b, inset).

The trial was terminated when the rat reached the goal, or after five minutes had elapsed. Upon reaching the goal, rats were confined there and given one Cheerio after a delay of approximately five seconds. This delay of reward procedure ensured that the rat could not locate the goal platform by its odour. The maze was also cleaned between trials to eliminate odour cues.

Two trial protocols were used in this study (see Testing schedules, and Extended Data Fig. 2 for further discussion of different schedules). Unoperated controls had three trials per day on the first two days and six trials per day for the next eleven days (72 trials in total). Operated controls were given four trials per day for seventeen days (68 trials in total). The number of choices per trial varied between 2 and 37 according to the rat's success rate, with a median of 5 choices per trial.

In the lesion study, the experimenter was blinded to lesion status.

Statistics. Differences in learning rates between the different control groups (unoperated versus sham hippocampal versus sham medial entorhinal cortex) were assessed using a two-way mixed ANOVA (Fig. 2). Two-way repeated measures ANOVAs were used to compare the performance of unoperated controls on trials with different values of angle β (Fig. 3a), with post hoc Bonferroni testing for pairwise comparisons (Fig. 3a, inset), and to test for an interaction between angle α ('ahead' choices versus 'behind' choices) and distance in trials in which choices consisted of adjacent platforms (Fig. 3d). One-way repeated measures ANOVAs were used to test for a relationship between performance and distance, and performance and angle α , in the unoperated control group (Fig. 3b–c). Performance on 'ahead' choices (angle $\alpha < 90^\circ$) versus 'behind' choices (angle $\alpha > 90^\circ$) was compared using a paired *t*-test, and a one-sample *t*-test was used to determine whether performance on 'behind' choices was significantly better than chance. A multiple regression analysis was conducted to evaluate the contributions of the three maze factors (angle α , angle β and distance) to performance in the unoperated control group (see main text and Extended Data Table 1).

Differences in learning curves were assessed in rats with hippocampal lesions and control rats using a two-way mixed ANOVA (Fig. 4a). Two-way mixed ANOVAs were also used to ascertain the effect of each maze variable (angle α , angle β and distance) on performance, and their potential interaction with lesion status (Fig. 4b–d). A Spearman's correlation was used to correlate lesion extent with performance (Extended Data Fig. 4). Differences in latencies between rats with hippocampal lesions and control rats over time were tested using a two-way mixed ANOVA, and a two-way mixed ANOVA was used to investigate the relationship between latencies, whether a choice was correct or incorrect and lesion status.

For Figs 2, 3a–c, 4 and Extended Data Fig. 5, data points indicate the mean score of one group of rats over a specified number of trials. In all figures, error bars indicate the s.e.m. For all statistical tests, data were tested to ensure they met the necessary assumptions before proceeding to analysis.

Histology. On completion of behavioural testing, rats were euthanized under anaesthesia and the brain fixed via intracardial perfusion with saline and 4% paraformaldehyde. Horizontal sections ($40 \mu\text{m}$) of the brains of rats with hippocampal lesions and of operated control rats with sham hippocampal lesions were mounted on gelatinized slides and stained with cresyl violet acetate. Slides were imaged using a Axio Scan.Z1 (Zeiss). Lesion volume was quantified by a blinded observer via manual tracing of the hippocampus on every fourth section, using ImageJ. The remaining volume of hippocampal tissue in rats with lesions was expressed as a percentage in relation to the measured volume of a typical operated control rat with a sham hippocampal lesion.

Testing schedules. Testing schedules were designed to meet a number of criteria. First, for each choice the rat was not offered a platform it had just occupied, with the exception of one instance, to eliminate the strategy of avoiding a platform that had immediately previously been occupied. Second, to prevent the task being solved using an egocentric strategy, correct choices were selected so that there was an approximately equal number of choices that required the rat to turn left (or anticlockwise), when facing the goal, as there were right (or clockwise) turns. Third, the starting platform changed between trials and these eight or nine platforms were distributed approximately equally around the maze.

The design of the maze enables a large number of different spatial navigation schedules to be run, from those most sensitive to hippocampal function to those least so. At the most sensitive end of the spectrum are schedules in which no trial is ever repeated, which rules out the possibility that the task can be solved by non-hippocampal dependent guidance or stimulus–response strategies. The animal must learn to approach a location in space from any starting position and any direction. At the other end of the spectrum are schedules that can easily be learned using a guidance strategy, such as always approaching a distal cue or a stimulus–response routine (for example, a sequence of body turns, such as left–right alternations). A guidance-biased training protocol might involve always starting the animal from the same location and offering the same sequence of choices, one of which always leads directly towards the goal (for an example, see ref. 19). Intermediate between these two schedules are ones that restrict the types

of choices available to the animal (for example, to choices between platforms at three different angles), which we have used in our first study to identify the factors contributing to successful performance.

Schedule 1. The schedule used in the unoperated control group was specifically designed to investigate the effect of the three metrics (angle α , angle β and distance) on task performance. In this experiment, the nine unoperated control rats undertook three different types of trial, which we named 'A β 60', 'A β 120' and 'A β 180'. In each trial type, the smallest angle between the two choice platforms offered (angle β) was fixed at 60° (A β 60), 120° (A β 120) or 180° (A β 180) degrees. The values of angle α were selected to ensure that for any given distance from the goal there were a range of values of angle α for each choice. This maximized the number of choices with unique combinations of distance, angle α and angle β , which enabled us to collect a dataset with 50 such unique combinations. Rats were tested in groups of three. Each rat completed six trials per day, which consisted of two trials of each type. In the testing schedule, trial type was staggered to control for the effect of experience on performance. Start platforms were also rotated among eight different locations and the combinations of start platforms and trial type were counterbalanced; Extended Data Fig. 2a illustrates the three trial types used in testing schedule 1.

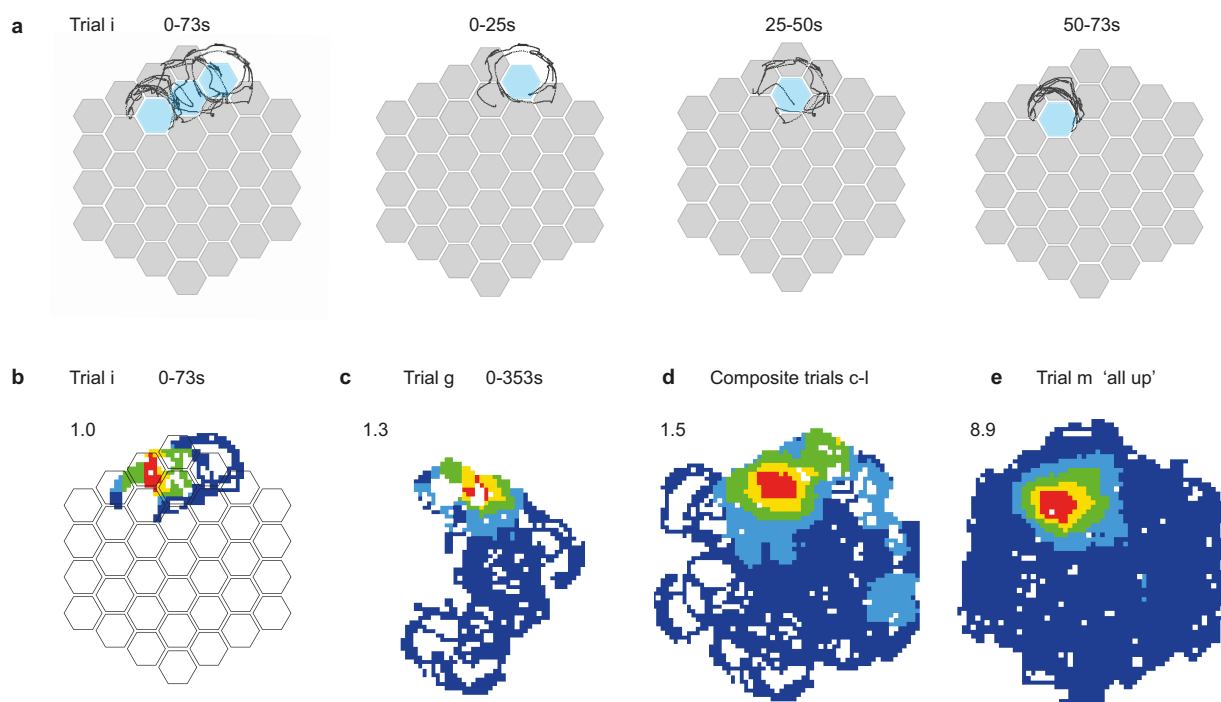
Schedule 2. When testing the rats with lesions, we designed a protocol that enabled us to examine the correlation between lesion size and task performance. In order to make a meaningful comparison between individual subjects with lesions, it was important to ensure that all rats experienced trials of the same difficulty in

the same temporal order. This was particularly important given the rapid rate at which rats learn the honeycomb maze task—as an illustration of this, on the first trial, control rats did not perform significantly better than rats with hippocampal lesions, but significantly outperformed them by the end of the first day of testing. For the lesion study, we therefore designed a schedule that contained choices with all possible values of angle β , to enable direct comparisons between rats with lesions and control rats on a day-by-day basis, and among rats with lesions of different sizes. Nine start platforms were used; this protocol is illustrated in Extended Data Fig. 2b.

Code availability. The custom software, written in LabView, used to run the maze trials is available from the corresponding author upon request.

Data availability. All relevant data are included within the paper and its Extended Data. Additional information including trial protocols is included in the Methods and Extended Data Fig. 2. The original and analysed datasets generated during the current study and any further methodological details required are available from the corresponding author upon request. The maze is designed to be as simple as possible, so that individual laboratories with access to a machine shop can build it. All detailed methods and materials on honeycomb maze building and control will be made available by the authors upon request, and we are currently searching for a manufacturer who can build and provide the mazes at a reasonable cost.

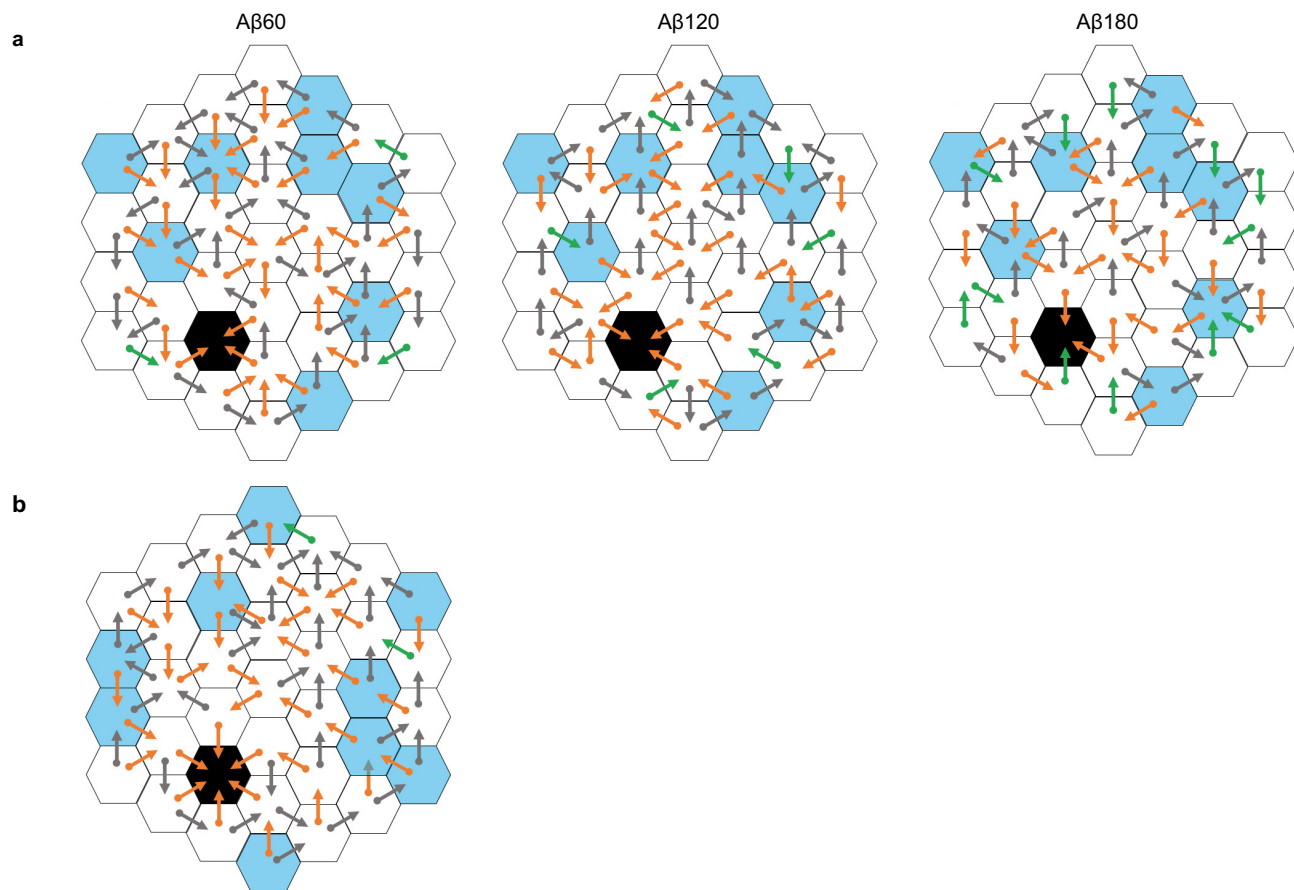
19. Eichenbaum, H., Stewart, C. & Morris, R. G. Hippocampal representation in place learning. *J. Neurosci.* **10**, 3531–3542 (1990).



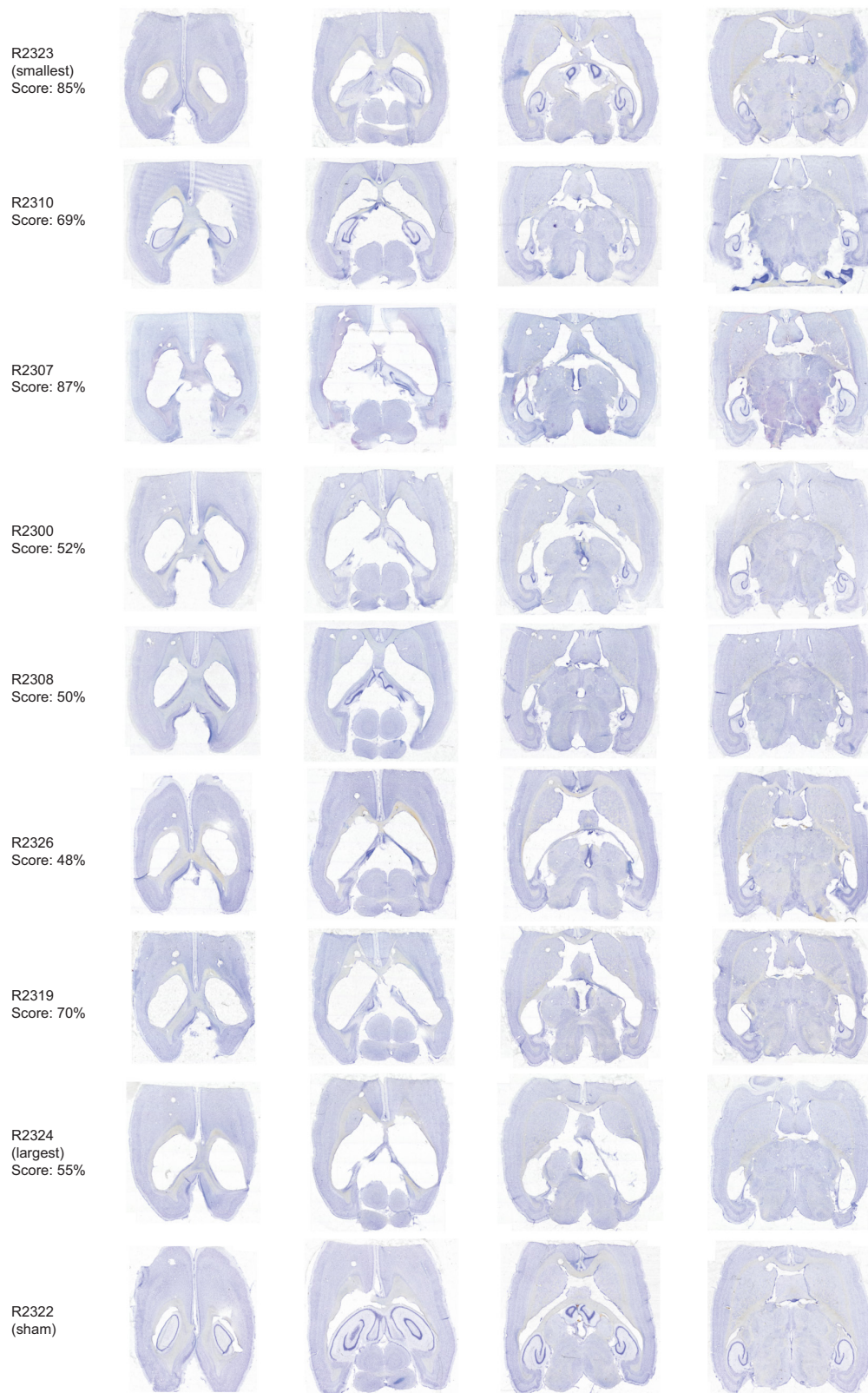
Extended Data Figure 1 | Place cell recorded on the honeycomb maze.

A single place cell recorded during navigation on the honeycomb maze. **a**, Behaviour (black line) from a single trial (trial *i*) in which the rat was offered two consecutive choices (left, 0–73 s) to go from the start platform (second left, 0–25 s), to an intermediate platform (third left, 25–50 s) to the goal (right, 50–73 s). The program detained the rat on each platform for 20 s before the two choice platforms were raised or, in goal, the food presented. Non-chosen platforms not shown. As the rat waited on each platform, it sampled the immediate environment by circling the perimeter of the platform. **b**, The firing of a place cell during this trial (trial *i*);

maximum rate in red shown top left. **c**, Rate map for the same cell on a separate trial (trial *g*) from a different start platform. **d**, Composite rate map from ten trials (trials *c–l*), each from a different starting location in which the rat took a different path to the goal. **e**, Firing rate map of the same cell when all platforms were raised and the rat foraged for food over a period of 20 min (trial *m*). In this cell, the firing fields during navigation trials (**d**) and during the foraging condition (**e**) were similar (spatial correlation = 0.77). Not all place cells displayed this profile, and others (not shown) fired in a different location(s) during the navigation trials from that seen in the foraging condition (that is, remapped).

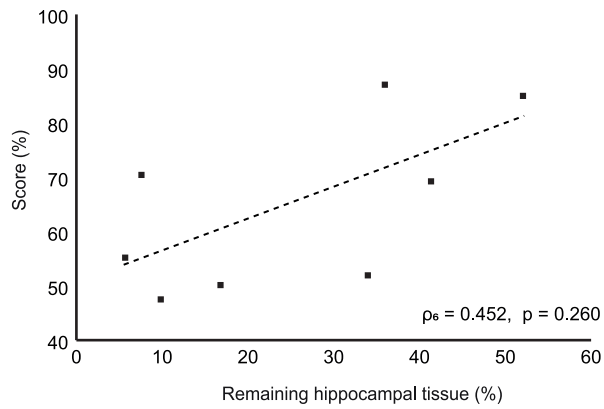


Extended Data Figure 2 | Protocols used on the honeycomb maze. a, Schedule 1 trial protocols for A β 60 (left), A β 120 (middle) and A β 180 (right) trials. **b,** Schedule 2 protocol. For **a** and **b**: goal platform, black; start platforms, blue; orange vectors, correct choices; grey vectors, incorrect choices; green vectors, 'forced' choices. See Methods for more detail.

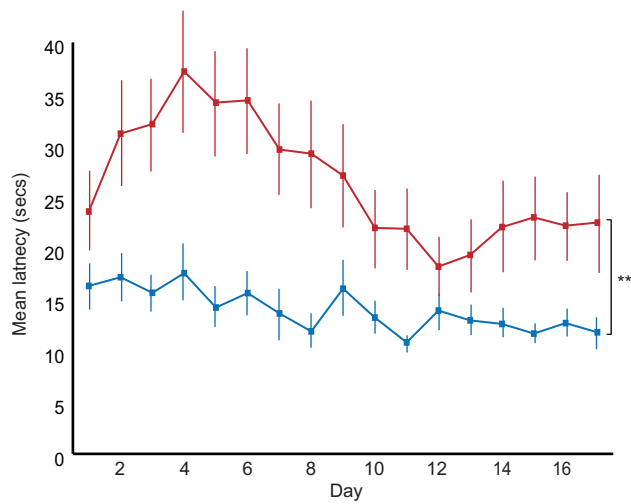


Extended Data Figure 3 | Histology of brains from rats with hippocampal lesions. Representative sections from brains of rats with hippocampal lesions, and one operated control with a sham hippocampal

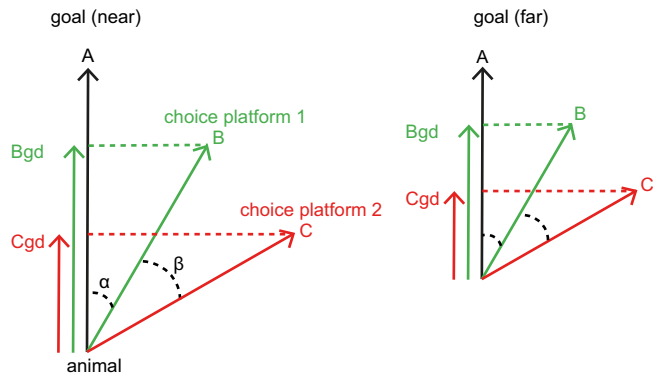
lesion (R2322), alongside mean performance scores on the honeycomb maze. Subjects are arranged in order of increasing lesion size. Horizontal sections (40 μ m) stained with cresyl violet.



Extended Data Figure 4 | Correlation between hippocampal volume and performance. Correlation between remaining hippocampal volume and performance on the spatial navigation task on the honeycomb maze in eight rats with hippocampal lesions ($n = 8$ rats, $\rho_s = 0.452$, $P = 0.260$; Spearman's correlation).



Extended Data Figure 5 | Rats with hippocampal lesions have longer latencies. Rats with hippocampal lesions (red, $n = 8$) have longer latencies than operated controls with sham hippocampal lesions (blue, $n = 8$) ($F_{1,14} = 11.103$, $P = 0.005$). Latencies also changed as a function of experience (day) ($F_{16,224} = 5.612$, $P < 0.001$) with a significant day \times lesion interaction ($F_{16,224} = 2.464$, $P = 0.002$, two-way mixed ANOVA). ** $P < 0.005$. Error bars indicate s.e.m.



Extended Data Figure 6 | Vector-based navigation schema. Left, The hippocampus represents a goal-direction vector pointing from the rat to the goal (A), which decreases as the rat is farther from the goal (right). The navigation system computes the projection of each choice platform vector (B , C) onto the goal-direction vector (inner product, B_{gd} , C_{gd}) and selects the larger of the two. This choice is easier with increased angle between choices (angle β) and consequent increased difference in the magnitudes of their projection vectors. The projection vector of the preferred platform, B_{gd} , is the output of the system that competes with other potential solutions to the problem (for example, choose between the leftmost or northmost platform).

Extended Data Table 1 | Summary of multiple regression analysis

Variable	B	SE _B	β	p value
Intercept	82.647	4.062		
Angle α	0.068	0.024	0.138	0.005
Angle β	- 1.862	0.930	- 0.099	0.046
Distance	- 0.091	0.029	- 0.157	0.002

A multiple regression analysis was undertaken to establish the contributions of angle α , angle β and distance to maze performance. Data were taken from nine unoperated control rats and performance assessed across all choice types. *B*, unstandardized regression coefficient; SE_B, standard error of the coefficient; β , standardized coefficient.

Extended Data Table 2 | Coordinates of the sites of hippocampal lesions, and the volume of ibotenic acid used in rats with lesions

Site	AP (mm)	ML (mm)	DV (mm)	IBO (μ l)
1	- 2.4	1.0	3.0	0.05
2	- 3.0	1.4	2.1	0.05
3	- 3.0	1.4	2.9	0.05
4	- 3.0	3.0	2.7	0.10
5	- 4.0	2.6	1.8	0.05
6	- 4.0	2.6	2.8	0.05
7	- 4.0	3.7	2.7	0.10
8	- 4.9	4.6	6.5	0.05
9	- 4.9	4.1	3.5	0.05
10	- 4.9	4.1	7.2	0.10
11	- 5.9	4.3	3.9	0.10
12	- 5.9	5.1	4.5	0.08
13	- 5.9	5.1	5.3	0.08
14	- 5.9	5.1	6.1	0.08

AP, anteroposterior with respect to bregma; ML, mediolateral with respect to bregma; DV, dorsoventral with respect to the brain surface; IBO, ibotenic acid.

Extended Data Table 3 | Coordinates of the sites of the sham medial entorhinal lesions

Site	ML (mm)	DV (mm)	AP coordinates
1	4.6	1.7	Angle: 22° along
2	4.6	2.2	AP axis pointing
3	4.6	2.7	rostrally
4	4.6	3.2	
5	4.6	3.7	Site: close to
6	4.6	4.2	transverse sinus
7	4.6	4.7	(without inflicting
8	4.6	5.2	damage)

AP, anteroposterior; ML, mediolateral with respect to lambda; DV, dorsoventral with respect to the brain surface. The glass micropipette was angled at 22° from vertical, pointing rostrally along the antero-posterior axis.

Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells

Joana Carrelha^{1,2}, Yiran Meng^{1,2}, Laura M. Kettle^{3,4}, Tiago C. Luis^{1,2}, Ruggiero Norfo^{1,2}, Verónica Alcolea^{1,2}, Francesca Grasso^{4,5}, Adriana Gambardella², Amit Grover², Kari Högstrand^{3,4}, Allegra M. Lord^{3,4}, Alejandra Sanjuan-Pla^{2†}, Petter S. Woll^{4,5}, Claus Nerlov^{2*} and Jacobsen Sten Eirik W.^{1,2,3,4,5*}

Rare multipotent hematopoietic stem cells (HSCs) in adult bone marrow (BM) with extensive self-renewal potential possess the ability to efficiently replenish all myeloid and lymphoid blood cells¹, securing long-term multilineage reconstitution following physiological and clinical challenges, including chemotherapy and hematopoietic transplantations^{2–4}. HSC transplantation remains the only curative treatment for many hematological malignancies, but inefficient blood-lineage replenishment remains a major cause of morbidity and mortality^{5,6}. Single cell transplantation has uncovered considerable heterogeneity among reconstituting HSCs^{7–11}, supported by findings in unperturbed hematopoiesis^{2–4,12} and suggested to reflect different propensities for lineage-fate decisions by distinct myeloid-, lymphoid- and platelet-biased HSCs^{7–10,13}. Other studies suggested that such lineage bias might reflect generation within the phenotypic HSC compartment of unipotent or oligopotent self-renewing progenitors, and implicated uncoupling of the defining HSC properties of self-renewal and multipotency^{11,14}. Here, highly sensitive tracking of progenitors and mature cells of the megakaryocyte/platelet, erythroid, myeloid, B and T cell lineages produced from singly transplanted HSCs revealed a highly organized, predictable and stable framework for lineage-restricted fates of long-term self-renewing HSCs. Most notably, a distinct class of HSCs adopts a fate towards effective and stable replenishment of a megakaryocyte/platelet-lineage tree but not other blood cell lineages, despite sustained multipotency, whereas no HSCs contribute exclusively to any other single blood-cell lineage. Single multipotent HSCs can also fully restrict towards simultaneous replenishment of megakaryocyte, erythroid and myeloid lineages without executing their sustained lymphoid lineage potential. Genetic lineage tracing supports an important role of platelet-biased HSCs also in unperturbed adult hematopoiesis. These findings uncover a limited repertoire of distinct HSC subsets, defined by a predictable and hierarchical propensity to adopt a fate towards replenishment of a restricted set of blood lineages, prior to loss of self-renewal and multipotency.

Previous studies of HSC lineage-bias transplanted CD45.2 HSCs into CD45.1 recipients^{7–9}, allowing tracking of myeloid (M) and B- and T-lymphoid progeny, but not replenishment of mature erythrocytes (E) and platelets (P), since they lack CD45 expression. Since recent studies also tracking E and/or P suggested distinct replenishment of these lineages^{10,11}, we purified Lin[–]Scal⁺c-Kit⁺CD34⁺CD150⁺CD48[–] (LSK34⁺150⁺48[–]) cells, containing most if not all HSCs^{15–17}, from BM of adult mice expressing enhanced green fluorescent protein (EGFP) from the *Gata1*-promoter (*Gata1*-EGFP)¹⁸ and tdTomato from the von Willebrand-factor promoter (*Vwf*-Tomato; Extended Data Fig. 1) to enrich platelet-primed *Vwf*^{pos} HSCs¹⁰. BM-ablated CD45.1 mice were

transplanted with a single CD45.2 *Vwf*-Tomato^{mid-high} (*Vwf*-Tomato^{pos}) LSK34⁺150⁺48[–] cell in competition with CD45.1 BM-cells. Through highly specific and sensitive analysis of large numbers of peripheral blood (PB) cells, contributions >0.01% to each blood lineage could reliably be detected (Methods). Almost 40% of transplanted single cells contributed to ≥0.1% of at least one lineage at 16–18 weeks post-transplantation, representing a candidate self-renewing long-term (LT)-HSC (Fig. 1a; Extended Data Fig. 2a). While approximately 50% replenished all 5 lineages (P, E, M, B and T), the other half showed no contribution (≤0.01%) to one or more lineages at all analysis time points, despite other lineages being robustly and stably replenished (Fig. 1b–d; Extended Data Fig. 2, 3), establishing a stringent definition of lineage-restricted reconstitution (Methods). If blood lineage-restricted reconstitution patterns by HSCs were randomly distributed, numerous lineage-restriction combinations would be expected, but only a few closely related patterns were observed in more than 1,000 single HSC transplanted mice. Most notably, 10% of long-term reconstituting *Vwf*-Tomato^{pos} LSK34⁺150⁺48[–] cells robustly and stably reconstituted platelets but no other lineages at any time-point (Fig. 1b–d; Extended Data Fig. 2b,d,e, 3a–b). No case was observed in which any other blood lineage was exclusively reconstituted (Fig. 1c–d; Extended Data Fig. 2e). In fact, the only additional stable lineage-restricted reconstitutions observed were P+E, P+E+M, and P+E+M+B, meaning that the only lineage invariably reconstituted was platelets (Fig. 1c,d; Extended Data Fig. 2c,d,e, 3c–e). Notably, also PE- and PEM-restricted replenishment consistently showed a strong P-bias (Fig. 1c,e; Extended Data Fig. 2d,f), although P-output was gradually reduced in mice with PEMB- versus PEM- versus P-restricted replenishment (Fig. 1c; Extended Data Fig. 2d; PEMB>PEM>P-restricted, $p < 0.05$ all time points; one-sided Wilcoxon-Mann-Whitney test). Parallel analysis of spleens confirmed the lineage-restricted reconstitution (Extended Data Fig. 3g–h). Previous studies used c-Kit-deficient (*W⁴¹/W⁴¹*) recipients and/or competitor cells^{7,8,10}, which might not only provide competitive advantage for transplanted wild-type HSCs but potentially also for development of blood cell lineages. Herein, almost identical lineage-restricted patterns, at similar frequencies (no significant differences), were observed whether single *Vwf*-Tomato^{pos} LSK34⁺150⁺48[–] cells were transplanted in competition with c-Kit-deficient (*W⁴¹/W⁴¹*; Fig. 1d) or wild-type (Extended Data Fig. 2e) BM-cells. In contrast, lineage-bias was affected by type of competitor, with a consistent tendency for enhanced lymphoid output in the presence of WT competitor cells (Fig. 1e, Extended Fig. 2f–g). This is compatible with the intrinsic propensity for distinct HSCs to adopt lineage-restricted fates (lineage-restriction) being more resistant to extrinsic influences, than to different lineage biases.

¹Haematopoietic Stem Cell Laboratory, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, UK. ²MRC Molecular Haematology Unit, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford OX3 9DS, UK. ³Department of Cell and Molecular Biology, Wallenberg Institute for Regenerative Medicine, Karolinska Institutet, Stockholm SE-171 77, Sweden. ⁴Department of Medicine Huddinge, Center for Hematology and Regenerative Medicine, Karolinska Institutet, Stockholm SE-171 77, Sweden. ⁵Karolinska University Hospital, Stockholm SE-171 77, Sweden. [†]Present address: Hematology Research Group, IIS La Fe, Valencia 46026, Spain.

*These authors contributed equally to this work.

The most primitive LSK150⁺48⁻ HSCs lack expression of CD229¹⁷. LSK150^{hi}48⁻CD229⁻ cells were enriched for *Vwf*-Tomato expression (Fig. 1f, Extended Fig. 4a), and transplanted single LSK34⁻150^{hi}48⁻CD229^{low/neg} cells were enriched for P-biased reconstitution when compared to LSK34⁻150^{hi}48⁻CD229^{hi} cells which were enriched for long-term lymphoid-biased reconstitution (Fig. 1g-h). No further distinct enrichment for *Vwf*-Tomato expression was observed based on CD41 expression (Extended Fig. 4b). Therefore, CD229 could partially replace the need for a transgenic *Vwf*-reporter to enrich platelet-biased HSCs.

Based on a different strategy for isolation and tracking of cells within the HSC-compartment, previous studies proposed the existence therein of lineage-restricted, oligopotent or unipotent, repopulating progenitors, including P-restricted progenitors, downstream of multipotent LT-HSCs¹¹. They reconstituted platelets much less efficiently and more transiently than multipotent LT-HSCs, suggesting a much more restricted self-renewal potential, confirmed by few sustaining platelets in secondary recipients¹¹. The herein long-term platelet-restricted replenishment by *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ HSCs was far more robust and stable (more than 50-fold higher at 16-18 weeks; Fig. 1c; Extended Data Fig. 2b,d, 3a,b), clearly distinguishing them from described platelet-restricted progenitors. In support of P-restriction already taking place within the *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ LT-HSC compartment, long-term stable P-restricted reconstitution by single *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ HSCs was invariably accompanied by robust long-term reconstitution of LSK150⁺48⁻ HSCs (mean 39%), virtually all being *Vwf*-Tomato^{pos} (mean 93%; Fig. 2a; Extended Data Fig. 5a). In contrast, in long-term multilineage reconstituted mice most reconstituted LSK150⁺48⁻ cells were *Vwf*-Tomato^{neg} (mean 7% *Vwf*-Tomato^{pos}; Fig. 2b; Extended Data Fig. 5b).

In mice with P-restricted and PEM-restricted reconstitution, BM and thymus lacked reconstitution of progenitors¹⁹⁻²¹ for lineages absent in PB, whereas robust progenitor-replenishment for reconstituted lineages was observed (Fig. 2a,c; Extended Data Fig. 5a,c). Subsets of multipotent progenitors (MPP) have been suggested to contain heterogeneous populations of lineage-restricted/biased progenitors in addition to true MPPs^{11,17,22,23}. LSKFlt3⁻CD150⁺CD48⁺ (150⁺48⁺) MPPs contain potent P-biased/restricted progenitors²³, LSKFlt3^{high}CD150⁺CD48⁺ lymphoid-primed MPPs (LMPPs)^{24,25}, LSKFlt3⁻CD150⁺CD48⁺ (150⁺48⁺) MPPs myeloid progenitor activity, and LSKFlt3⁻CD150⁺CD48⁻ (150⁺48⁻) MPPs multi-potent MPPs with considerable reconstitution potential^{17,23}. All MPP-compartments were robustly replenished in multilineage reconstituted mice, whereas long-term P-restricted reconstitution showed high chimerism for 150⁺48⁺ MPPs, with no detectable replenishment of 150⁺48⁺ MPPs, LMPPs or 150⁺48⁻ MPPs. In PEM-restricted reconstitution 150⁺48⁺ and 150⁺48⁻ MPPs were replenished whereas LMPPs and 150⁺48⁻ MPPs remained undetectable in most mice (Fig. 2a-c; Extended Data Fig. 5). Whereas most 150⁺48⁺ cells in P-restricted reconstituted mice were *Vwf*-Tomato^{pos} (mean 87%), a small fraction (mean 3%) were *Vwf*-Tomato^{pos} in multilineage-reconstituted mice, suggesting that *Vwf*-Tomato^{pos} P-restricted reconstituting LSK150⁺48⁻ cells replenish a distinct platelet-primed subset of 150⁺48⁺ MPPs. This detailed BM-analysis of mice exclusively and robustly reconstituted with PB-platelets by a single platelet-primed LSK34⁻150⁺48⁻ HSC tentatively identifies a platelet-restricted lineage tree, apparently independent of most implicated MPP-subsets.

Our detailed analysis in mice with P-restricted reconstitution failed to provide evidence for progenitor reconstitution beyond those potentially dedicated to the megakaryocyte-platelet lineage. This is compatible with long-term P-restricted LSK34⁻150⁺48⁻ reconstituting cells representing unipotent progenitors with fully P-restricted potential as proposed in previous studies¹¹, although also compatible with self-renewing multipotent LSK34⁻150⁺48⁻ platelet-primed LT-HSCs adopting a lineage-restricted fate exclusively toward P-restricted replenishment prior to loss of multipotency. To distinguish between

these two fundamentally distinct possibilities, we sorted CD45.2⁺ LSK cells from mice with robust and stable P-restricted reconstitution four months after transplantation of a single *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ HSC, and assessed potential to produce granulocytes, monocytes/macrophages and T-lymphocytes *in vitro*²⁶. Remarkably, for all investigated mice with P-restricted reconstitution, definitive morphological, FACS and gene expression data unequivocally established that single *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ LT-HSCs exclusively producing platelets *in vivo*, also possess extensive granulocyte, monocyte/macrophage and T-lymphoid potential *in vitro* (Fig. 2d-g, Extended Data Fig. 6). Therefore, *Vwf*-Tomato^{pos} LSK34⁻150⁺48⁻ cells executing potent P-restricted LT-reconstituting activity represent self-renewing multipotent HSCs that exclusively adopt a megakaryocyte-platelet lineage fate upon transplantation *in vivo*.

Upon extended analysis for up to 44 weeks, high and dominant P-reconstitution was sustained in all mice, although in 5 out of 13 mice with P-restricted reconstitution at 16-18 weeks, very low-level erythrocyte and/or myeloid reconstitution was also observed (Fig. 3a-b). Self-renewal potential of P-restricted *Vwf*-Tomato^{pos} HSCs, and whether P-restricted reconstitution potential is intrinsically programmed, was assessed in secondary transplantations. We consistently observed sustained high levels of P-restricted secondary reconstitution, although in several secondary recipients very low levels of erythroid, myeloid and even lymphocyte reconstitution was detected (Fig. 3c), providing further support for the multipotency of single *Vwf*-Tomato^{pos} HSCs exclusively adopting a fate towards platelet-replenishment in primary recipients. Secondary recipients also displayed highly selective, robust and stable reconstitution of the *Vwf*-Tomato^{pos} LSK150⁺48⁻ HSC, *Vwf*-Tomato^{pos} 150⁺48⁺ MPP, and MkP progenitor axis, as in corresponding primary recipients (Fig. 3d-e; Extended Data Fig. 7a). These experiments established that single P- or PE-restricted LT-HSCs, sustain self-renewal and distinct lineage-restricted fates, in secondary recipients.

To more unequivocally establish whether P- or PE-restricted self-renewing LT-HSCs sustain their multipotency, we purified *Vwf*-Tomato^{pos} LSK150⁺48⁻ cells from the BM of mice with robust P- or PE-restricted reconstitution 4 months after transplantation of a single *Vwf*-Tomato^{pos} LSK150⁺48⁻ HSC, and transplanted these into secondary recipients (Extended Data Fig. 7b). Following an additional 4 months, sustaining P- and PE-restricted reconstitution (Fig. 3f), we established that the original single *Vwf*-Tomato^{pos} LSK150⁺48⁻ HSC transplanted had sustained its multipotency (Fig. 3g), despite continuing to strictly adopt a P- or PE-restricted fate *in vivo*. Since these secondary transplantations were performed using fresh BM competitor cells, rather than whole BM cells from the primary recipients as in most studies^{7,8}, they also rule out that the sustained lineage-restricted fate is explained by a distinct and constant composition of competing HSCs and progenitors.

The physiological relevance, if any, of platelet-biased HSCs in steady-state unperturbed hematopoiesis remains unclear, as recent HSC fate mapping studies were not designed to assess this, including clonal tracking by barcoding in which platelets were not assessed since they do not harbour DNA used to retrieve barcodes^{2-4,12}. A *Vwf*-CreERT2 mouse line selectively labelling a fraction of LSK34⁻48⁻ cells with high CD150 expression (Fig. 3h), selectively, robustly and increasingly with time labelled platelets in adult mice (Fig. 3i). Up to 2 months after a short Tamoxifen-pulse, almost exclusively platelets were labelled, subsequently along with a small fraction of erythrocytes and myeloid cells, but virtually no lymphocytes, compatible with the labelling and significant activity of herein identified P- and PEM-biased HSCs in steady-state hematopoiesis.

Single cell index-sorting²⁷ was performed to assess lineage reconstitution as a function of *Vwf*-Tomato expression in LSK34⁻150⁺48⁻ BM cells (Fig. 4a). *Vwf*-Tomato^{mid-high} LSK34⁻150⁺48⁻ fractions showed a higher frequency of repopulating cells than *Vwf*-Tomato^{neg} cells (Fig. 4b). In agreement with the proposed existence of lymphoid-biased HSCs^{7-9,13,28}, a large fraction of mice transplanted with single *Vwf*-Tomato^{neg} LSK34⁻150⁺48⁻ cells showed robust long-term

lymphoid-biased or even lymphoid-restricted PB-reconstitution (Fig. 4c–d), independently of level of CD150 expression (Extended Data Fig. 8a–c). However, in contrast to P-, PE or PEM-restricted reconstituting LT-HSCs (which never showed evidence of replenishment of any other lineages), lymphoid-restricted/biased long-term PB reconstitution was always preceded by transient multi-lineage reconstitution (Fig. 4d–e; Extended Data Fig. 8d,e), more compatible with persistence of long-lived lymphocytes²⁹ or lymphoid progenitors produced by short-term multipotent HSCs than the ongoing activity of a lymphoid-biased HSC. In further support of this, and in agreement with previous studies,^{7,8} long-term lymphoid-biased or lymphoid-restricted reconstitution was not associated with detectable replenishment of the LSK34⁺150⁺48⁺ HSC compartment (Extended Data Fig. 9), and also accompanied by reduced PB and progenitor reconstitution in secondary recipients (Fig. 4f, Extended Data Fig. 10a–c). Thus, while *Vwf*-Tomato^{neg} LSK34⁺150⁺48⁺ short-term repopulating cells might play an important role in replenishment of lymphocytes, detailed PB kinetics and HSC and progenitor reconstitution analysis, fail to support programming of lymphoid-fate restriction at the LT-HSC-level, as here demonstrated for P-, PE-, PEM- and PEMB-restricted HSCs.

Excluding lymphoid-biased/restricted and other short-term repopulating cells, distinct patterns of PB-lineage reconstitution were observed among LSK34⁺150⁺48⁺ LT-HSCs with different *Vwf*-Tomato levels (Fig. 4g–k; Extended Data Fig. 10d). Robust long-term multilineage reconstitution was observed in all fractions, with a clear tendency towards increasing platelet-bias with increasing levels of *Vwf*-Tomato expression. P-restricted reconstitution was exclusively a property of *Vwf*-Tomato^{pos} and never seen with *Vwf*-Tomato^{neg} cells. *Vwf*-Tomato^{mid} and *Vwf*-Tomato^{low} P-restricted reconstitution was at lower levels than from *Vwf*-Tomato^{high} cells (Extended Data Fig. 8a). Therefore, multipotent HSCs robustly adopting a P-restricted fate predominantly have a platelet-primed *Vwf*-Tomato^{high} phenotype.

In conclusion, highly sensitive tracking of PB-replenishment of P-, E-, M-, B- and T-cell lineages, and their progenitors, in mice transplanted with single cells establish the existence of four distinct and closely related stages of self-renewing HSCs in adult BM that stably adopt lineage-restricted fates despite remaining multipotent, therefore establishing that a uni-lineage P-restricted fate determination can occur already in self-renewing LT-HSCs with sustained multipotency. This suggests that the programme required for multipotency might also be required for the self-renewal capacity of LT-HSCs adopting a stable maintenance of a P-restricted fate. PEMB-, PEM-, PE- and P-restricted multipotent HSCs all reside in the LSK34⁺150⁺48⁺ *Vwf*-Tomato^{pos} platelet-primed LT-HSC compartment¹⁰, and typically show a distinct and stable platelet-bias in their reconstitution. Although this is compatible with a hierarchy of multipotent HSCs gradually restricting their fate towards exclusive megakaryocyte-platelet production, the hierarchical or non-hierarchical relationship between multipotent HSCs which adopt distinct lineage-restricted fates, remains to be established. Findings in mouse^{10,11} and human³⁰ showing that platelet levels increase and decrease more rapidly than other blood-cell lineages suggest that their progenitors are short-lived, and must therefore be continuously replenished at a high rate, which could explain the need for HSCs dedicated to platelet-replenishment. Despite having transplanted more than 1,000 mice, we found no evidence in support of the existence of LT-HSCs with a fate restricted to any other blood cell lineages. The sustained and serially transplantable P-restricted blood lineage replenishment suggests that the P-restricted fate of multipotent *Vwf*-Tomato^{high} HSCs is intrinsically programmed, most likely through epigenetic mechanisms. While the LT-HSCs adopting a P-restricted fate sustains multipotency, it remains to be established at what subsequent stage in the progenitor hierarchy established by P-restricted HSCs, multipotency is eventually lost¹¹. Unraveling these mechanisms, and developing approaches to expand P-restricted HSCs, is likely to help ameliorate extensive needs for platelet transfusions following chemotherapy and BM transplantation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 January; accepted 18 December 2017.

Published online 3 January 2018.

- Osawa, M., Hanada, K.-I., Hamada, H. & Nakauchi, H. Long-term lymphohematopoietic reconstitution by a single CD34-low/negative hematopoietic stem cell. *Science* **273**, 242–245 (1996).
- Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014).
- Busch, K. *et al.* Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature* **518**, 542–546 (2015).
- Sawai, C. M. *et al.* Hematopoietic stem cells are the major source of multilineage hematopoiesis in adult animals. *Immunity* **45**, 597–609 (2016).
- Seggewiss, R. & Einsele, H. Immune reconstitution after allogeneic transplantation and expanding options for immunomodulation: an update. *Blood* **115**, 3861–3868 (2010).
- Pineault, N. & Boyer, L. Cellular-based therapies to prevent or reduce thrombocytopenia. *Transfusion* **51 Suppl 4**, 72S–81S (2011).
- Müller-Sieburg, C. E., Cho, R. H., Thoman, M., Adkins, B. & Sieburg, H. B. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood* **100**, 1302–1309 (2002).
- Dykstra, B. *et al.* Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell* **1**, 218–229 (2007).
- Challen, G. A., Boles, N. C., Chambers, S. M. & Goodell, M. A. Distinct hematopoietic stem cell subtypes are differentially regulated by TGF- β 1. *Cell Stem Cell* **6**, 265–278 (2010).
- Sanjuan-Pla, A. *et al.* Platelet-biased stem cells reside at the apex of the haematopoietic stem-cell hierarchy. *Nature* **502**, 232–236 (2013).
- Yamamoto, R. *et al.* Clonal analysis unveils self-renewing lineage-restricted progenitors generated directly from hematopoietic stem cells. *Cell* **154**, 1112–1126 (2013).
- Pei, W. *et al.* Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature* **548**, 456–460 (2017).
- Wang, J. *et al.* Per2 induction limits lymphoid-biased haematopoietic stem cells and lymphopoiesis in the context of DNA damage and ageing. *Nat Cell Biol* **18**, 480–490 (2016).
- Haas, S. *et al.* Inflammation-induced emergency megakaryopoiesis driven by hematopoietic stem cell-like megakaryocyte progenitors. *Cell Stem Cell* **17**, 422–434 (2015).
- Kiel, M. J., Yilmaz, O. H., Iwashita, T., Terhorst, C. & Morrison, S. J. SLAM family receptors distinguish hematopoietic stem and progenitor cells and reveal endothelial niches for stem cells. *Cell* **121**, 1109–1121 (2005).
- Yilmaz, O. H., Kiel, M. J. & Morrison, S. J. SLAM family markers are conserved among hematopoietic stem cells from old and reconstituted mice and markedly increase their purity. *Blood* **107**, 924–930 (2006).
- Oguro, H., Ding, L. & Morrison, S. J. SLAM family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell* **13**, 102–116 (2013).
- Drissen, R. *et al.* Distinct myeloid progenitor-differentiation pathways identified through single-cell RNA sequencing. *Nat Immunol* **17**, 666–676 (2016).
- Nagasawa, T. Microenvironmental niches in the bone marrow required for B-cell development. *Nat Rev Immunol* **6**, 107–116 (2006).
- Bhandoola, A., von Boehmer, H., Petrie, H. T. & Zúñiga-Pflücker, J. C. Commitment and developmental potential of extrathymic and intrathymic T cell precursors: plenty to choose from. *Immunity* **26**, 678–689 (2007).
- Pronk, C. J. *et al.* Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell* **1**, 428–442 (2007).
- Wilson, A. *et al.* Hematopoietic stem cells reversibly switch from dormancy to self-renewal during homeostasis and repair. *Cell* **135**, 1118–1129 (2008).
- Pietras, E. M. *et al.* Functionally distinct subsets of lineage-biased multipotent progenitors control blood production in normal and regenerative conditions. *Cell Stem Cell* **17**, 35–46 (2015).
- Mansson, R. *et al.* Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. *Immunity* **26**, 407–419 (2007).
- Adolfsson, J. *et al.* Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential: a revised road map for adult blood lineage commitment. *Cell* **121**, 295–306 (2005).
- Böiers, C. *et al.* Lymphomyeloid contribution of an immune-restricted progenitor emerging prior to definitive hematopoietic stem cells. *Cell Stem Cell* **13**, 535–548 (2013).
- Wilson, N. K. *et al.* Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. *Cell Stem Cell* **16**, 712–724 (2015).
- Eaves, A. C. Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood* **125**, 2605–2613 (2015).
- Sprent, J. & Tough, D. F. Lymphocyte life-span and memory. *Science* **265**, 1395–1400 (1994).

30. Vellenga, E. *et al.* Autologous peripheral blood stem cell transplantation in patients with relapsed lymphoma results in accelerated haematopoietic reconstitution, improved quality of life and cost reduction compared with bone marrow transplantation: the Hovon 22 study. *British Journal of Haematology* **114**, 319–326 (2001).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank A.J. Mead, D. Atkinson, A. Giustacchini and N. Ashley for expert assistance with the Fluidigm array platform [WIMM Single Cell Core Facility is supported by the MRC MHU (MC_UU_12009), the Oxford Single Cell Biology Consortium (MR/M00919X/1) and the WT-ISSF (097813/Z/11/B#) funding, the WIMM Strategic Alliance awards G0902418 and MC_UU_12025]; P. Sopp and S-A. Clark for expert flow cytometry technical support and cell sorting services [WIMM FACS Core Facility is supported by the MRC HIU, MRC MHU (MC_UU_12009), NIHR Oxford BRC and the John Fell Fund (131/030 and 101/517), the EPA fund (CF182 and CF170) and by WIMM Strategic Alliance awards (G0902418 and MC_UU_12025)]; the Biomedical Services at University of Oxford for animal technical support; the EMBL Monterotondo Gene Expression Service and Transgenic Core Facility for generating the *Vwf*-tdTomato BAC and the corresponding transgenic mouse line; N. Iscove (Ontario Cancer Institute, University Health Network) for *W⁴¹/W⁴¹* mice; A. Cumano (Institut Pasteur) for OP9-DL1 stromal cells; R. Drissen (Weatherall Institute of Molecular Medicine) for helpful discussions; S. Duarte and H. Boukarabila (Weatherall Institute of Molecular Medicine) for assistance with preliminary phase of the studies; A. Hillen (Karolinska Institutet),

B. Wu and T. Bouriez-Jones (Weatherall Institute of Molecular Medicine) for technical assistance. This work was supported by Marie Curie Early Stage Researcher Fellowship (J.C.), the Medical Research Council UK (G0801073 and MC_UU_12009/5 to S.E.W.J. and G0701761, G0900892 and MC_UU_12009/7 to C.N.), the Swedish Research Council (S.E.W.J.), the Knut och Alice Wallenberg Foundation (WIRM; S.E.W.J.), the Tobias Foundation (S.E.W.J.), StratRegen KI (S.E.W.J.), and a BBSRC Project Grant (BB/M024350/1; C.N.).

Author Contributions S.E.W.J. and C.N. conceptualized the research, with input from A.S-P and J.C. S.E.W.J., C.N., J.C., Y.M., L.M.K. and P.S.W. designed the experiments and analysed the data. J.C. and Y.M. performed all experiments except fate mapping, with assistance from T.C.L., A.Ga. and A.Gr. (single cell transplantations), L.M.K., R.N. and V.A. (peripheral blood reconstitution analysis), and F.G. (CD229/CD41 analysis). L.M.K. performed fate mapping experiments with assistance from K.H. and A.M.L. S.E.W.J., C.N. and J.C. wrote the manuscript, which was subsequently reviewed and approved by all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to S.E.W.J. (sten.jacobsen@imm.ox.ac.uk), sten.eirik.jacobsen@ki.se or C.N. (claus.nerlov@imm.ox.ac.uk).

Reviewer Information *Nature* thanks E. Laurenti, S. Morrison and the other anonymous reviewer(s) for their contribution to the peer review of this work.

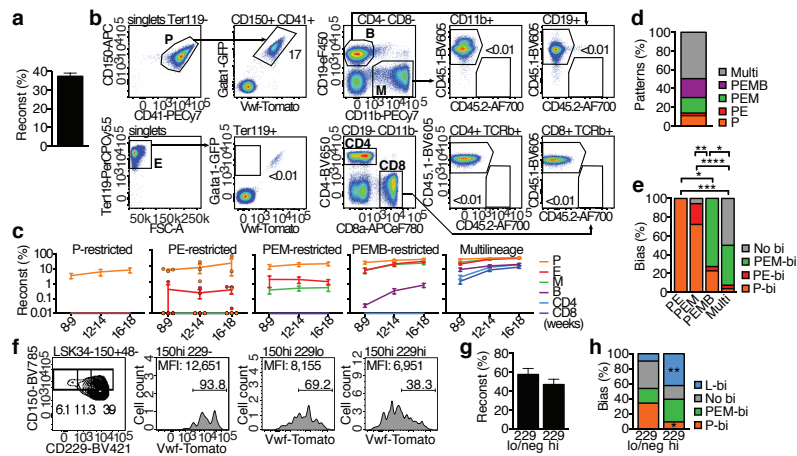


Figure 1 | Stable long-term lineage-restricted reconstitution patterns by single *Vwf-Tomato*^{pos} LSK34⁺150⁺48⁻ cells. **a**, Reconstituted mice (mean±s.e.m.; ≥1 lineage ≥0.1% at 16–18wks) transplanted with single *Vwf-Tomato*^{pos} LSK34⁺150⁺48⁻ HSC and *W⁴¹/W⁴¹* BM support. n = 292 mice, 15 experiments. **b**, Analysis of platelet, erythroid-, myeloid-, B- and T-cell reconstitution in P-restricted patterns 16–18wks post-transplantation, representative of >30 mice in >40 single-cell transplantation experiments. **c**, Reconstitution kinetics (mean±s.e.m.). P-restricted, n = 12 mice; PE-restricted, n = 3; PEM-restricted, n = 18; PEMB-restricted, n = 22; Multilineage, n = 54. **d**, Distribution of lineage-restricted reconstitution patterns (lineage positive if >0.01% in

≥1 analysis point; n = 109 mice). **e**, Distribution of lineage-bias within restriction patterns in **d** (all plotted in **c**). P-bi frequency: Multi vs PEMB *p = 0.019, PEMB vs PE *p = 0.024, **p = 0.004, ***p = 0.0003, ****p < 0.001. **f**, *Vwf-Tomato* expression in LSK34⁺150⁺48⁻ cells with different CD229 expression. Plots representative of 6 mice in 4 experiments. Percentage of parent gate and median fluorescence intensity (MFI). **g**, Reconstituted mice (mean±s.e.m.) transplanted with single CD229^{low/neg} (n = 71 mice) or CD229^{high} (n = 70) LSK34⁺150⁺48⁻ cells in 3 experiments. **h**, Distribution of lineage-bias by CD229^{low/neg} (n = 41 mice) or CD229^{high} (n = 33) single cells. *p = 0.013; **p = 0.002. Statistical comparisons: two-tailed Fisher's exact test (95% CI).

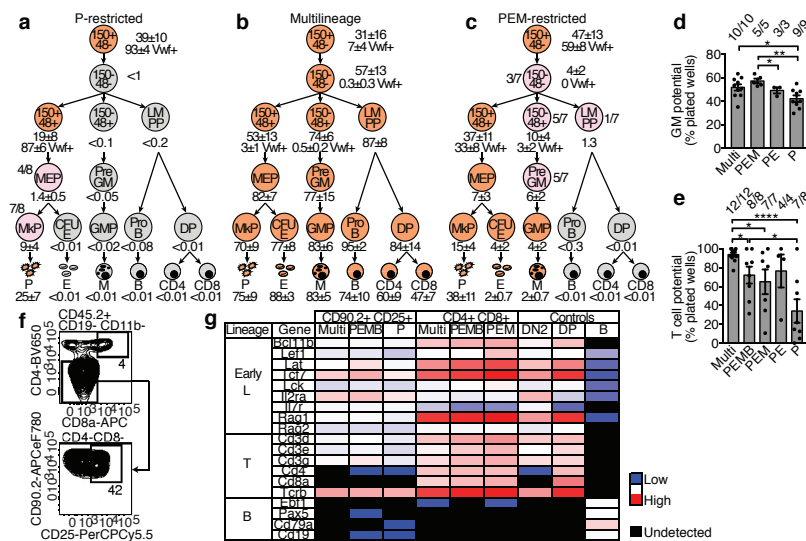


Figure 2 | Single reconstituting LSK34⁺150⁺48⁺ cells establish lineage-restricted hematopoietic hierarchies but remain multipotent.

a-c, Reconstitution percentages (mean \pm s.e.m.) of HSPCs 16-44wks post-transplantation in P-restricted (**a**, n = 8 mice, except Pro-B n = 5), Multilineage (**b**, n = 6, except Pro-B n = 3) and PEM-restricted mice (**c**, n = 7, except Pro-B n = 4 and DP n = 6) in 12 experiments. Orange: all mice positive. Grey: no positive mice. Pink: positive/negative mice, frequency and mean \pm s.e.m. of positives shown. **d**, GM *in vitro* potential of donor-derived LSK cells (mean \pm s.e.m. per group, 1 cell/well, 60-240 wells/mouse, 20 experiments). Multi, n = 10 mice; PEM-restricted, n = 5; PE-restricted n = 3; P-restricted, n = 9. Fraction of positive mice above bars. P vs Multi *p = 0.023; PE vs PEM *p = 0.047; **p = 0.003. **e**, T-cell *in vitro* potential, donor-derived LSK cells (mean \pm s.e.m., 10 cells/well,

7-72 wells/mouse, 21 experiments). Multi, n = 12 mice; PEMB-restricted, n = 8; PEM-restricted, n = 7; PE-restricted, n = 4; P-restricted, n = 8. Fraction of positive mice above bars. P vs PEMB *p = 0.026; PEM vs Multi and PEMB vs Multi *p = 0.011; ****p < 0.0001. f, FACS-profile of T-cell progenitors generated *in vitro*, representative of 21 experiments. g, Gene expression heatmap of *in vitro* T-cell progenitors. CD90.2⁺CD25⁺ samples: Multi, n = 3 mice; PEMB-restricted, n = 3; P-restricted, n = 3. CD4⁺CD8⁺ samples: Multi, n = 4 mice; PEMB-restricted, n = 2; PEM-restricted, n = 1. 2-8 wells/mouse, 5 experiments. Controls: DN2 (CD90.2⁺CD25⁺) thymocytes, n = 4 mice; DP (CD4⁺CD8⁺) thymocytes, n = 5 mice; CD19⁺CD11b⁺ spleen B-cells, n = 4 mice. Mean Ct values per mouse group, normalized to mean Ct of *Hprt1/B2m*. Statistical comparison of means: two-tailed t-test (95% CI).

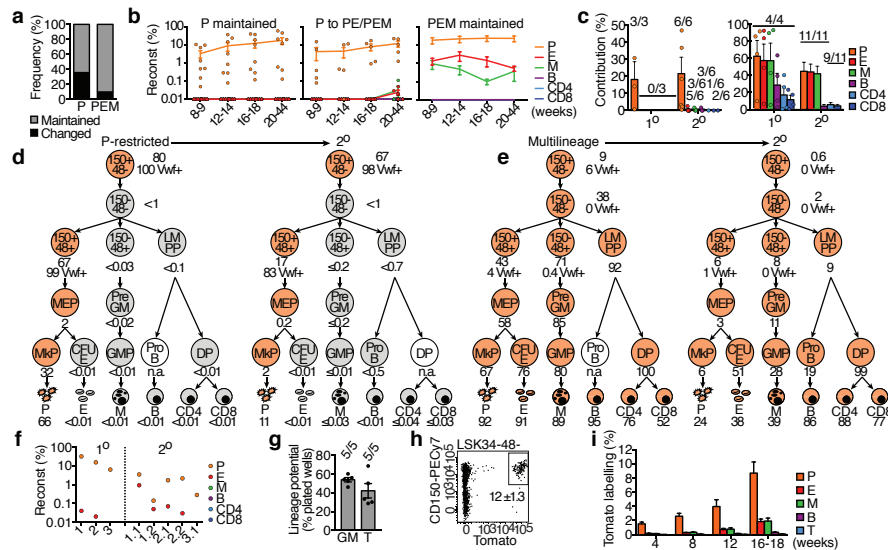


Figure 3 | Long-term persistence of lineage-restricted reconstitution patterns by multipotent HSCs. **a**, Frequency of strict long-term maintenance of P- and PEM-restricted reconstitution by single *Vwf*-Tomato^{pos} HSCs, observed at all analysis points for a total 20-44wks (P-restricted, n = 14 mice; PEM-restricted, n = 10; 12 experiments). **b**, Reconstitution kinetics (mean ± s.e.m.) in P- and PEM-restricted mice maintained (n = 9 each) or changed (n = 5) analysed ≥20wks post-transplantation. **c**, Reconstitution (mean ± s.e.m. of positive mice) 16-18wks post-transplantation in primary and secondary recipients. P-restricted, n = 3 primary with 6 secondary, 1-3/donor (left; 2 experiments). Multilineage, n = 4 primary with 11 secondary, 1-4/donor (right; 4 experiments). Frequency positive mice above bars. **d-e**, Reconstitution percentages (mean ± s.e.m.) of HSPCs 44wks post-transplantation of a single HSC. P-restricted (**d**) and Multilineage (**e**), and for each a corresponding secondary recipient 26wks post-transplantation. Orange: all mice positive. Grey: no positive mice. White: not analysed.

f, PE- and P-restricted reconstitution, 16-18wks post-transplantation, in primary recipients of a single *Vwf*-Tomato^{pos} HSC and their secondary recipients of donor-derived *Vwf*-Tomato^{pos} HSCs (3 experiments). **g**, GM (1 cell/well, 19-37 wells/mouse) and T cell (10 cells/well, 6-19 wells/mouse) *in vitro* potentials of donor-derived LSK cells from secondary recipients in **f**, 19-21wks post-transplantation (mean ± s.e.m., n = 5 mice). Fraction of positive mice above bars. **h**, HSC labelling 2-3wks post-Tamoxifen treatment of *Vwf*-CreERT2/R26-Tomato mice (mean ± s.e.m. % of LSK34-150+48-, n = 3). **i**, Blood labelling (mean ± s.e.m. post-Tamoxifen treatment of *Vwf*-CreERT2/R26-Tomato mice (5 experiments; n = 14 mice at 4 and 8wks; n = 11 at 12 and 16-18wks). Statistical comparison of means with two-tailed t-test: 4 and 8wks, P vs each lineage, all p < 0.0001; 12wks, P vs E p = 0.003, P vs M p = 0.004, P vs B p = 0.0007, P vs T p = 0.0005; 16-18wks, P vs E p = 0.0004, P vs M p = 0.0005, P vs B and P vs T p < 0.0001.

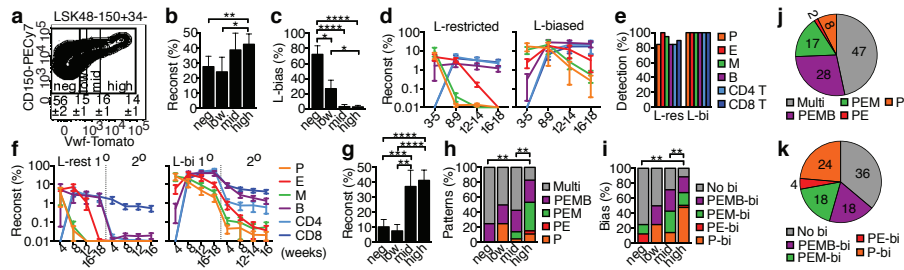


Figure 4 | Lymphoid-biased reconstitution independently of long-term reconstituting HSCs. **a**, *Vwf*-Tomato gating in post-sort analysis of index-sorted LSK34⁺150⁺48⁺ cells. Mean percentages \pm s.e.m., $n = 14$ mice, 14 experiments. **b**, Reconstituted mice (mean \pm s.e.m.; any lineage $\geq 0.1\%$ at 16–18 wks) transplanted with a single HSC and WT BM support in 14 experiments. *Vwf*^{neg}, $n = 107$ mice; *Vwf*^{low}, $n = 43$; *Vwf*^{mid}, $n = 68$; *Vwf*^{high}, $n = 174$. * $p = 0.023$; ** $p = 0.007$. **c**, Frequency of lymphoid-bias in **b**. *Vwf*^{neg} vs *Vwf*^{low} * $p = 0.030$; *Vwf*^{low} vs *Vwf*^{high} * $p = 0.021$; **** $p < 0.0001$. **d**, Reconstitution kinetics (mean \pm s.e.m.). L-restricted: lymphoid $\geq 0.1\%$ and absence of non-lymphoid lineages 16–18 wks ($n = 19$ mice). L-biased: lymphoid reconstitution $\geq 0.1\%$ and ≥ 1 non-lymphoid lineage at 16–18 wks ($n = 5$). **e**, Frequency of detection of PB lineages ($\geq 0.1\%$ reconstitution) in ≥ 1 analysis point (mice in **d**). **f**, Reconstitution

kinetics (mean \pm s.e.m.) in secondary recipients of BM from L-restricted (left; $n = 7$ primary mice and 21 secondary mice in 3 experiments, 2–4 secondary/donor) and L-biased mice (right; $n = 6$ primary and 15 secondary in 2 experiments, 2–3 secondary/donor). **g**, Reconstituted mice (mean \pm s.e.m) with M and/or P $\geq 0.1\%$ at 16–18 wks, excluding L-biased reconstitution (same mice in **b**). ** $p = 0.0098$; *** $p = 0.0001$; **** $p < 0.0001$. **h**, Distribution of restriction patterns in mice in **g**. Ratio of patterns with vs without lymphoid output: *Vwf*^{neg} vs *Vwf*^{high} ** $p = 0.006$; *Vwf*^{mid} vs *Vwf*^{high} ** $p = 0.002$. **i**, Distribution of lineage-bias in mice in **g**. P-bi frequency: *Vwf*^{neg} vs *Vwf*^{high} ** $p = 0.009$; *Vwf*^{mid} vs *Vwf*^{high} ** $p = 0.006$. **j–k**, Distribution of lineage restriction (**j**) and bias (**k**) in total LSK34⁺150⁺48⁺ LT-HSCs (combining data in **a, g, h, i**). Statistical comparisons: two-tailed Fisher's exact test (95% CI).

METHODS

Animals. All mice were bred and maintained, and all experimental procedures were performed, in accordance with UK Home Office regulations. All experiments were approved by the Oxford University Clinical Medicine Ethical Review Committee.

All mouse strains were backcrossed for more than six generations onto a C57BL/6J background. Both male and female adult mice (7 to 18 weeks old) were used in experiments. When multiple experimental groups were analysed, mice were allocated so that each group was evenly matched for age range and frequency of mice of each sex. Index-sorted single cells were transplanted randomly within each cohort of recipient mice. No statistical methods were used to predetermine the experimental sample size. The frequency of reconstitution patterns observed in preliminary single cell transplantation experiments was used to estimate the minimum number of single-cell transplanted mice required in order to observe ≥ 5 mice with each of the peripheral blood reconstitution patterns identified. Additional mice were transplanted with single cells in order to obtain enough biological replicates of relevant reconstitution patterns for all lines of experiments. Investigators were not blinded to experimental group allocation or when assessing experimental outcomes.

B6.SJL-Ptprca⁺ Pepc^b/BoyJ mice (CD45.1) were used as primary and secondary transplantation recipients, and in some experiments (CD45.1 x C57BL/6J)F1 mice (CD45.1/2) were also used as secondary recipients. As specified, wild type CD45.1 or *Ki⁺W⁻⁴¹/W⁻⁴¹-Gpi1^{a/a}* mice³¹ with CD45.1 allotype (*W⁴¹/W⁴¹*) were used as donors of unfractionated bone marrow (BM) support cells in primary competitive transplantation experiments. The *Vwf*-tdTomato transgene was generated by bacterial artificial chromosome (BAC) recombinant engineering at the EMBL Genome Engineering Core Facility. To validate the hematopoietic expression pattern of *Vwf*-tdTomato, its co-expression with *Vwf*-EGFP was analysed after being crossed to similarly constructed and previously published *Vwf*-EGFP mice¹⁰. To produce single cell donors for primary transplantations in which both platelet and erythroid reconstitution could be evaluated, *Vwf*-tdTomato mice were crossed to the previously published *Gata1*-EGFP mice¹⁸, generating *Vwf*-tdTomato/*Gata1*-EGFP double reporter mice, heterozygous for both transgenic alleles and expressing the CD45.2 allotype. A *Vwf*-CreERT2 line was generated by Cyagen Biosciences by knock-in of an IRES-CreERT2-P2A-EGFP cassette into 3'UTR of the endogenous *Vwf* gene in C57BL/6 mice. For fate mapping experiments, this *Vwf*-CreERT2 line was crossed to B6;129S6-Gt(Rosa)^{tm9(CAG-tdTomato)Hze/J} (R26R-tdTomato). EGFP expression was not detectable in these mice.

Single cell transplantations. Single cell transplantations were performed as previously published¹⁰, using a FACSAriaII, FACSAriaIII or FACSAria Fusion cell sorter (BD Biosciences). BM cell suspensions were prepared by crushing leg, sternum and spine bones into PBS (Gibco) with 5% fetal calf serum (FCS, Gibco) and 2mM ethylenediaminetetraacetic acid (EDTA, Sigma-Aldrich). Single phenotypically-defined Lin⁻Scal⁺c-Kit⁺CD34⁺CD150⁺CD48⁻ BM hematopoietic stem cells (HSCs, LSK34⁺150⁺48⁻) were sorted from *Vwf*-tdTomato/*Gata1*-EGFP (CD45.2) adult mice (antibody details in Supplementary Table 1) into Iscove's Modified Dulbecco's Medium (IMDM, Gibco) with 20% BIT-9500 Serum Substitute (Stem Cell Technologies), 100U/mL Penicillin and 0.1mg/mL Streptomycin (100x Pen/Strep, PAA Laboratories), 2mM L-Glutamine (PAA Laboratories) and 0.1mM 2-Mercaptoethanol (Sigma-Aldrich). Single cells were deposited by an automated cell deposition unit (ACDU). Accurate deposition of single fluorescent beads was assessed before and after single cell sorting to ensure that a single cell was deposited into all wells. Single sorted HSCs were mixed with 1×10^6 *W⁴¹/W⁴¹* or 2×10^5 wild-type CD45.1 unfractionated BM support cells, and transplanted by intravenous injection into lethally irradiated (10 Gy) CD45.1 mice.

Single transplanted HSCs were sorted as *Vwf*-Tomato^{pos} (*Vwf*-Tomato^{mid-high}) or, in order to determine the significance of *Vwf*-tdTomato expression level, were index sorted and based on the fluorescence intensity of *Vwf*-tdTomato for each cell were grouped into *Vwf*^{neg}, *Vwf*^{low}, *Vwf*^{mid} or *Vwf*^{high} categories. The same indexing strategy was used to identify CD150^{high} and CD150^{low} transplanted single cells. FACS reanalysis confirmed that mean sort purity (defined as fulfilment of all sort gating criteria upon reanalysis) was 95%.

Secondary transplantations. In bulk BM secondary transplantation experiments, 5×10^6 or 30×10^6 unfractionated BM cells from single-cell reconstituted primary recipient mice were resuspended in PBS 1% FCS and transplanted by intravenous injection into lethally irradiated (10 Gy) CD45.1 or CD45.1/2 recipients. Secondary transplantations into lethally irradiated CD45.1 mice were also performed with donor-derived LSK34-150+48- HSCs sorted into the IMDM-based single-cell transplantation media described above, together with fresh CD45.1 BM support cells.

Peripheral blood reconstitution analysis. PB samples were collected from tail vein into lithium heparin-coated microvettes (Sarstedt). A small aliquot of

unfractionated PB was used for analysis of erythroid cells, separately or mixed with platelet solution. Platelets were separated by centrifugation of PB samples at 100xg for 10 minutes at room temperature. For leukocyte separation, samples were incubated 1:1 with 2% w/v Dextran (Sigma-Aldrich), for 20-30 minutes at 37°C. Excess erythrocytes were lysed from the leukocyte preparation through incubation with ammonium chloride solution (Stem Cell Technologies) for 2 minutes at room temperature, and leukocyte samples were pre-incubated with Fc-block prior to staining. Anti-mouse antibody staining for reconstitution analysis by flow cytometry was carried out for 15-20 minutes at 4°C in PBS 1% FCS 2mM EDTA (antibody details in Supplementary Table 1), and samples were analysed using LSRII, LSR Fortessa or LSR Fortessa X-20 flow cytometers (BD Biosciences).

Reconstituted erythroid cells were identified as Ter119⁺ *Vwf*-Tomato^{neg} *Gata1*-GFP⁺. Reconstituted platelets were identified as Ter119⁻CD150⁺CD41⁺ *Vwf*-Tomato^{pos} *Gata1*-GFP⁺. To investigate if aggregation between platelets derived from the transplanted single HSC and derived from the support cells might result in an overestimation of platelet reconstitution levels in transplanted mice, we mixed *Vwf*-Tomato^{pos} *Gata1*-GFP⁺ and wild-type PB to obtain platelet ratios of approximately 1:1 ($n = 6$), 1:4 ($n = 6$) and 1:10 ($n = 3$). The mixed PB samples were processed for flow cytometry as detailed above, and for the expected percentages of approximately 50%, 25% and 10% *Vwf*-Tomato^{pos} *Gata1*-GFP⁺ platelets we as predicted observed $54.1\% \pm 5.4\%$, $27.6\% \pm 3.9\%$ and $7.0\% \pm 1.0\%$, respectively. We also transplanted CD45.1 recipients with 20×10^6 unfractionated BM cells in which the ratio of *Vwf*-EGFP to *Vwf*-Tomato cells was 1:100 or 1:2 ($n = 3$ recipients for each). With the 1:100 ratio we observed $0.2\% \pm 0.09$ of Tomato+GFP+ aggregates among labeled platelets, and with the 1:50 ratio we observed $1.3\% \pm 0.39$, indicating that aggregation between platelets is not substantial.

Leukocyte cell populations in PB were defined as follows: myeloid cells, NK1.1⁻CD4⁻CD8⁻CD19⁻CD11b⁺; B cells, NK1.1⁻CD4⁻CD8⁻CD11b⁻CD19⁺; CD4 T cells, NK1.1⁻CD19⁻CD11b⁻CD8⁻CD4⁺TCR3⁺; CD8 T cells, NK1.1⁻CD19⁻CD11b⁻CD4⁻CD8⁺TCR3⁺. Contribution to leukocytes in PB and spleen by CD45.2⁺ transplanted single cells was identified by co-expression of CD45.2 (and lack of CD45.1 expression) and mature lineage markers (antibody details in Supplementary Table 1), and calculated based on the frequency of CD45.2⁺ cells among total CD45⁺ events for each lineage, excluding CD45 negative events and CD45.1⁺CD45.2⁺ double-positive artifacts.

PB reconstitution in primary transplant recipients was analysed by flow cytometry at 8-10, 12-14 and 16-18 weeks post-transplantation for all mice, and also at 3-5 weeks and 20-44 weeks in specific extended kinetics experiments. Mice were considered to be reconstituted by a potential long-term HSC when donor contribution to at least one PB lineage was $\geq 0.1\%$ at 16-18 weeks post-transplantation.

Reconstitution patterns generated by potential long-term HSCs were considered to be restricted when reconstitution of one or more PB lineages was consistently below the detection level ($\leq 0.01\%$; see detection thresholds section below) at all time points (at least 3) of PB analysis up to 18 weeks post-transplantation. Therefore, for a transplanted HSC to be classified as being P-restricted, no reconstitution of neither E, M, B or T cells was observed at any time point in PB up to 16-18 weeks. Likewise, patterns in which each PB lineage was detectable ($> 0.01\%$) in at least one time point up to 18 weeks post-transplantation were considered to be Multilineage. According to these criteria, four lineage-restricted patterns were found to be produced by *Vwf*^{mid-high} LSK150⁺48⁺34⁻ cells: PEMB-, PEM-, PE- and P-restricted. When transplanting *Vwf*^{neg} and *Vwf*^{low} cells, another relevant pattern was also found, and designated LM-restricted. This pattern was defined by reconstitution of lymphoid (B and/or T cells) and at least one, but not all, myelo-erythroid lineages at one or more analysis points (3-5, 8-10, 12-14, 16-18 weeks).

Lineage-bias was defined as a sustained threefold higher lineage output compared to the remaining PB lineages, observed at 12-14 as well as 16-18 weeks post-transplantation. Patterns in which no single lineage or combination of lineages was threefold higher than all other lineages were then considered to be without a specific bias ("no bias"), and so were the patterns with no concordance of lineage-bias observed at 12-14 weeks and 16-18 weeks. The L-bias category includes mice in which lymphoid lineage (B and/or T) reconstitution is threefold higher than platelets and myeloid cells, but not necessarily threefold higher than erythroid cells due to their slower decline after the HSC exhaustion that underlies L-bias output.

In initial experiments we observed that no primary recipients with $\leq 0.01\%$ reconstitution of any of the PB lineages at 8-10 weeks post-transplantation were reconstituted by a long-term HSC, since in these mice ($n = 26$) reconstitution did not reach $\geq 0.1\%$ for any lineage at 16-18 weeks. Therefore, in further primary transplantation experiments recipients with $\leq 0.01\%$ reconstitution of all lineages at ≥ 8 weeks post-transplantation were not analysed further. Secondary recipients were always analysed at 8-10, 12-14 and 16-18 weeks post-transplantation, and

also at 3–5 weeks if transplanted with BM cells from Lymphoid-restricted and Lymphoid-biased primary reconstituted recipients. Recipient mice that did not survive or had to be sacrificed before 16–18 weeks after primary or secondary transplantation were eliminated from the analysis (4.5% of a total of 528 mice). Secondary recipients without detectable reconstitution ($>0.01\%$) of any myelo-erythroid lineage (platelets, erythroid and myeloid cells) at 16–18 weeks post-transplantation and/or at the terminal analysis time point (18–28 weeks post-transplantation) were not considered to be HSC-reconstituted and therefore eliminated from secondary reconstitution analysis, except if being recipient of BM cells from Lymphoid-restricted and Lymphoid-biased primary reconstituted recipients.

Reconstitution analysis of hematopoietic stem and progenitor cells. Reconstitution analysis of lymphoid progenitors in thymus and hematopoietic stem and progenitor cell (HSPC) populations in BM was performed using LSRII, LSR Fortessa or LSR Fortessa X-20 flow cytometers (BD Biosciences). Except in the myeloid progenitor panel, which involves staining with fluorophore-conjugated CD16/32, BM, spleen and thymus cells were pre-incubated with Fc-block prior to anti-mouse antibody staining (antibody details in Supplementary Table 1). Reconstitution was calculated by quantifying the contribution of transplanted CD45.2⁺ cells to each cell population, except in the case of CFU-E in which *Gata1*-GFP⁺ cells were quantified instead due to the low expression level of CD45.1 and CD45.2 in these erythroid progenitors, which have been shown to be $\sim 100\%$ *Gata1*-GFP⁺ 18.

BM HSPC populations were defined as follows. HSC: Lin[−]Sca1⁺c-Kit⁺(LSK) Flt3[−]CD150⁺CD48[−]; 150⁺48⁺ MPP: LSKFlt3[−]CD150⁺CD48⁺; 150⁺48⁺ MPP: LSKFlt3[−]CD150⁺CD48[−]; LMPP: LSKFlt3^{high}CD150⁺CD48⁺; MkP: Lin[−]Sca1⁺c-Kit⁺(LK)CD150⁺CD41⁺; MEP: LKCD41[−]CD16/32[−]CD150⁺CD105[−]; CFU-E: LKCD41[−]CD16/32[−]CD150⁺CD105⁺; PreGM: LKCD41[−]CD16/32[−]CD150⁺CD105[−]; GMP: LKCD41[−]CD150⁺CD16/32⁺; Pro-B: Lin[−]B220⁺CD19⁺c-Kit⁺ (in some experiments further defined as IgM⁺). Thymus T cell progenitors were defined as Lin[−]CD4⁺CD8⁺. We observed that the size of the classically-defined LMPP population (25% highest Flt3⁺ cells within LSK)²⁵ varied significantly between donor-derived CD45.2⁺ LSK and support-derived CD45.1⁺ LSK in different reconstituted mice, and therefore an LMPP gate equivalent to the 25% highest Flt3-expressing LSK cells in 7–18 week old non-transplanted mice was used as described before²⁵.

Detection thresholds for reconstitution analysis by flow cytometry. PB from non-transplanted CD45.1 and CD45.1/2 mice was screened alongside experimental samples to evaluate the minimum threshold of reconstitution detection. In the CD45.1 and CD45.1/2 controls ($n \geq 10$ for each), $\geq 10,000$ events were recorded for each PB lineage. The detection threshold for *Vwf*-Tomato^{pos} *Gata1*-GFP⁺ platelets and *Vwf*-Tomato^{neg} *Gata1*-GFP⁺ erythroid cells was found to be 0.01% for both recipient types. The detection threshold for CD45.2 false positive events in the leukocyte lineages was found to also be 0.01% in CD45.1 recipients, and 0.03% in CD45.1/2 recipients. Reconstitution levels above these threshold levels were only considered reliable if also ≥ 5 donor-derived events could be recorded within a lineage gate. Reconstitution detection threshold for myeloid, B and T cells in spleen was also found to be 0.01%, based on analysis of CD45.1 controls ($n = 2$).

As in PB, minimum reconstitution detection thresholds for progenitors were based on HSPC analysis of non-transplanted CD45.1 and CD45.1/2 controls ($n \geq 4$ for each), and defined as follows for both recipient types. 150⁺48⁺ MPP and 150⁺48⁺ MPP, $>0.2\%$; HSC and LMPP, $>0.1\%$; PreGM, $>0.03\%$; MEP and Pro-B, $>0.02\%$; GMP, CFU-E and thymus DP, $>0.01\%$. The thresholds indicated for HSPC reconstitution were frequently higher than these minimum threshold values due to the rarity of some HSPC populations in transplanted mice. In some cases, as specified, the number of events that could be recorded by flow cytometry was too low to reach the minimum threshold levels, additionally because reconstitution levels above the minimum thresholds were only considered reliable if ≥ 5 donor-derived events could be recorded. When averaging reconstitution percentages across several biological replicates, mean detection thresholds were calculated based on mean number of events recorded for each HSPC population.

Analysis of *Vwf*-tdTomato/*Vwf*-EGFP mice. To validate the novel *Vwf*-Tomato reporter mouse line, co-expression of *Vwf*-Tomato and *Vwf*-GFP in *Vwf*-tdTomato/*Vwf*-EGFP double reporter mice was analysed in PB platelets and erythroid cells, as well as BM LSK34⁺150⁺48⁺, MEP, MkP, PreCFU-E, CFU-E, PreGM and GMP. PB and BM HSPC analysis was performed as detailed above, with the addition of a c-Kit/CD117 enrichment step for HSPCs. Briefly, 2×10^6 unfractionated BM cells were resuspended in 200 μ L PBS 5% FCS 2mM EDTA and incubated with 5 μ L anti-mouse CD117 MicroBeads (Miltenyi Biotec) for 20 minutes at 4 °C, then processed through MACS LS columns (Miltenyi Biotec) according to manufacturer instructions.

Fate mapping analysis of *Vwf*-CreERT2/R26R-tdTomato mice. The *Vwf*-CreERT2 mouse line was generated by Cyagen Biosciences. Cre recombinase

activation was induced in 8–18 weeks old *Vwf*-CreERT2/R26R-tdTomato mice by Tamoxifen administration. Tamoxifen powder (Sigma T5648) was dissolved in a sterile solution of 10% ethanol and 90% corn oil, for a final concentration of 20mg/mL. This solution was administered to the mice by oral gavage, 4mg per day for 5 consecutive days.

Peripheral blood from Tamoxifen-induced *Vwf*-CreERT2/R26R-tdTomato mice was processed similarly to that of transplanted mice. In order to control for potential aggregation of labelled and non-labelled cells, a *Gata1*-EGFP CD45.1 blood spike-in was introduced to experimental samples ($n = 14$) at two different time points for each sample. CD41 staining was included to exclude platelets adhering to white blood cells.

In vitro myeloid lineage potential assays and validation. Granulocyte-macrophage (GM) lineage potential was evaluated *in vitro* in X-VIVO15 liquid medium containing Gentamycin and L-Glutamine (Lonza) and supplemented with 10% FCS (GE Healthcare HyClone), 0.1mM 2-Mercaptoethanol (Sigma-Aldrich), and the following cytokines: 2ng/mL mouse stem cell factor (mSCF, PeproTech), 5ng/mL human FLT3 ligand (hFL, Immunex), 5ng/mL human thrombopoietin (hTPO, PeproTech), 5ng/mL mouse interleukin 3 (mIL-3, PeproTech), 10ng/mL human granulocyte colony-stimulating factor (hG-CSF, Neopogen) and 10ng/mL mouse GM colony-stimulating factor (mGM-CSF, Immunex). Donor-derived (CD45.2⁺) LSK cells were bulk sorted from single-cell transplanted CD45.1 mice and manually plated into 4 wells of round-bottom 96-well plates (Corning Costar) at a density of 25–50 cells per well, or into 60–240 wells of 60-well Terasaki microplates (Thermo Fisher Scientific) at a density of 1 cell per well. Culture confluency in Terasaki microplates was scored at day 10–14, and presence of mature myeloid cells was confirmed at day 13–18 by morphological analysis of cytospin slides (prepared as described below), and gene expression analysis by multiplex quantitative PCR (described below) of 4 wells per biological replicate (TaqMan probe details in Supplementary Table 2). Myeloid cells generated by 25–50 donor-derived LSK cells were also analysed by flow cytometry for co-expression of CD45.2, CD11b and Ly6G.

In vitro T cell lineage potential assay and validation. T cell lineage potential was evaluated *in vitro* in OP9-DL1 stromal co-cultures, as previously described³². GFP⁺ OP9-DL1 stromal cells (cell line provided by A. Cumano, Institut Pasteur) were maintained in adherent cultures in 75cm² flasks (Corning) with Opti-MEM Glutamax medium (Gibco) supplemented with 10% FCS, 0.1mM 2-Mercaptoethanol, 100U/mL Penicillin and 0.1mg/mL Streptomycin. Expected properties of OP9-DL1 cell line were validated: distinct morphology, GFP expression, and capacity to robustly promote T cell differentiation from hematopoietic stem and progenitor cells. OP9-DL1 cell stocks were regularly tested for mycoplasma contamination by PCR.

One or two days before sorting the cells of interest for the assay, adherent cells were trypsinized (Trypsin EDTA, PAA Laboratories) and monolayers of OP9-DL1 stromal cells were prepared by seeding $1.5\text{--}2 \times 10^3$ cells per well in flat-bottom 96-well plates (Corning Costar). Donor-derived (CD45.2⁺) LSK cells were bulk sorted and manually plated at a density of 10 cells per well into 7–72 wells of pre-prepared confluent monolayers ($\sim 80\%$) of OP9-DL1 stromal cells. Cultures were maintained for 4–5 weeks in culture medium supplemented with the following cytokines: 5ng/mL mSCF and 5ng/mL hFL during first 1–2 weeks of culture, then mSCF only; or 5ng/mL mSCF and 5ng/mL hFL throughout, with addition of 4–5ng/mL hIL-7 during final week. Cultures were analysed by flow cytometry for detection of CD45.2⁺CD90.2⁺CD25⁺ and CD45.2⁺CD4⁺CD8⁺ T cell progenitors using LSRII, LSR Fortessa or LSR Fortessa X-20 cytometers (antibody details in Supplementary Table 1). T cell gene expression analysis was performed by multiplex quantitative PCR (TaqMan probe details in Supplementary Table 2) on samples of 25–50 donor-derived T cell precursors sorted from 2–8 wells from each biological replicate using a FACSARIA Fusion sorter (BD Biosciences).

In vitro lineage potential assays for secondary recipients. Lethally-irradiated (10Gy) CD45.1 mice ($n = 3$ recipients per donor) were transplanted with donor-derived (CD45.2⁺) LSK34⁺150⁺48⁺ *Vwf*-Tomato^{pos} cells sorted at 18–26 weeks post-transplantation from primary recipients with PE- or P-restricted patterns ($n = 3$ donors). Each secondary recipient received 250–1,000 donor-derived sorted cells together with 2×10^5 wild-type CD45.1 support cells. 19–21 weeks after secondary transplantation, donor-derived LSK cells were sorted from the secondary recipients with maintained PE- or P-restricted patterns and tested in GM and T cell *in vitro* lineage potential assays as described in the previous section.

GM potential: bulk sorted CD45.2⁺ LSK cells were manually plated at a density of 1 cell per well into 20–37 wells of 60-well Terasaki microplates. Cell confluency per well was scored at day 13, and presence of mature myeloid cells was confirmed by morphological analysis.

T cell potential: bulk sorted CD45.2⁺ LSK cells were manually plated at a density of 10 cells per well into 6–19 wells of pre-prepared monolayers of OP9-DL1 stromal cells. Cultures were maintained for 4 weeks before flow cytometry analysis,

and identity of phenotypic CD45.2+ T cell progenitors was confirmed by gene expression analysis.

Cytospins and cell morphology analysis. Myeloid cultures grown *in vitro* were collected into PBS with 20% FCS (Gibco) and centrifuged onto polylysine-coated microscope slides (VWR) by a Shandon Cytospin 4 Cytocentrifuge (Thermo Scientific). Slides were stained with May Grünwald (Sigma-Aldrich) and Giemsa (Fluka) solutions according to manufacturer instructions, and mounted with Pertex (Cellpath). Morphology of stained cells was analysed with an Olympus BX60 upright compound microscope, images were captured with INFINITY imaging software (Lumenera) and edited with ImageJ software (NIH, public domain).

Multiplexed quantitative PCR. Multiplex quantitative PCR analysis was performed using the BioMark 48.48, 96.96 or 192.24 Dynamic Array platform (Fluidigm) and TaqMan Gene Expression Assays (Thermo Fisher Scientific) as previously described^{32,33}. Cells were sorted or cultures were transferred into PCR tubes containing 10 µL or 15 µL of RT/pre-amplification buffer, respectively. CellsDirect One-Step qRT-PCR kit (Invitrogen) was used for cDNA synthesis and pre-amplification of target genes. 10 µL of pre-amplification buffer consisted of 2.5 µL TaqMan assay mix containing all assays at 0.2x dilution, 5 µL CellsDirect 2x Reaction mix (Invitrogen), 1.2 µL CellsDirect RT/Taq mix (Invitrogen), 1.2 µL TE buffer and 0.1 µL SUPERase-In RNase Inhibitor (Ambion). Targeted cDNA pre-amplification was performed during 22 cycles and pre-amplified product was diluted 1:5 in TE buffer before processing with Dynamic Array protocol according to manufacturer instructions. Details of TaqMan Gene Expression Assays (Life Technologies) in Supplementary Table 2. Controls with no template or with no RT polymerase were also analysed in each experiment, as well as positive and negative control cell samples.

Heatmaps were generated in Microsoft Excel using data analysed by the Δ Ct method, using the mean of housekeeping genes Hypoxanthine Guanine Phosphoribosyltransferase 1 (*Hprt1*) and Beta-2-microglobulin gene (*B2m*) for

normalization. Only samples expressing *Hprt1*, *B2m*, and the pan-hematopoietic CD45 gene (*Ptprc*) were included for further analysis.

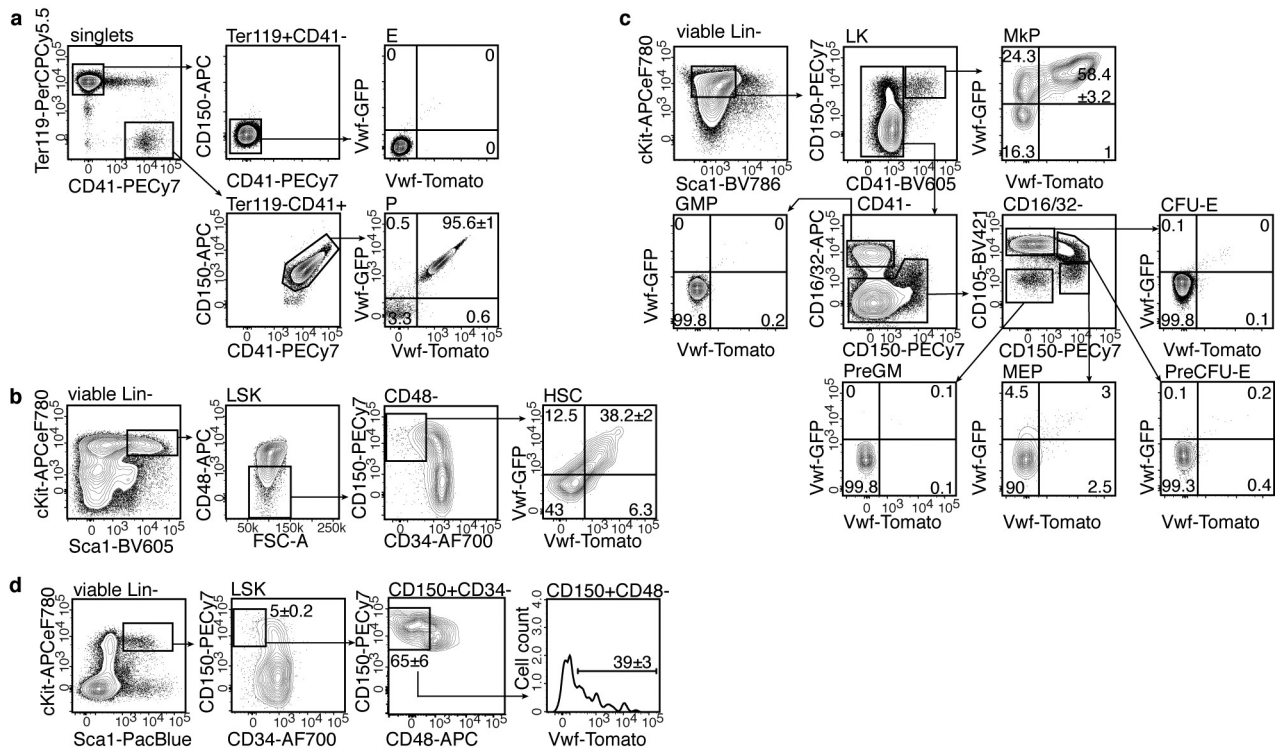
Statistics. Statistical comparisons of continuous data with assumed normal distribution (mean percentages) were performed with parametric unpaired two-tailed Student's *t*-test. Statistical comparisons of categorical data (frequencies) were performed with nonparametric two-tailed Fisher's exact test. Statistical comparison of platelet reconstitution levels between patterns was performed with one-sided Wilcoxon Mann-Whitney test (normal distribution not assumed based on Shapiro-Wilk normality test).

Only statistically significant differences ($p < 0.05$) were indicated in the figures. For all statistical comparisons: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Student's *t*-test and Fisher's exact test were performed with GraphPad Prism software and GrapPad QuickCalcs online tool: www.graphpad.com/quickcalcs. Wilcoxon-Mann-Whitney test was performed with online tool: ccb-compute2.cs.uni-saarland.de/wtest/?id=www/www-ccb/html/wtest³⁴.

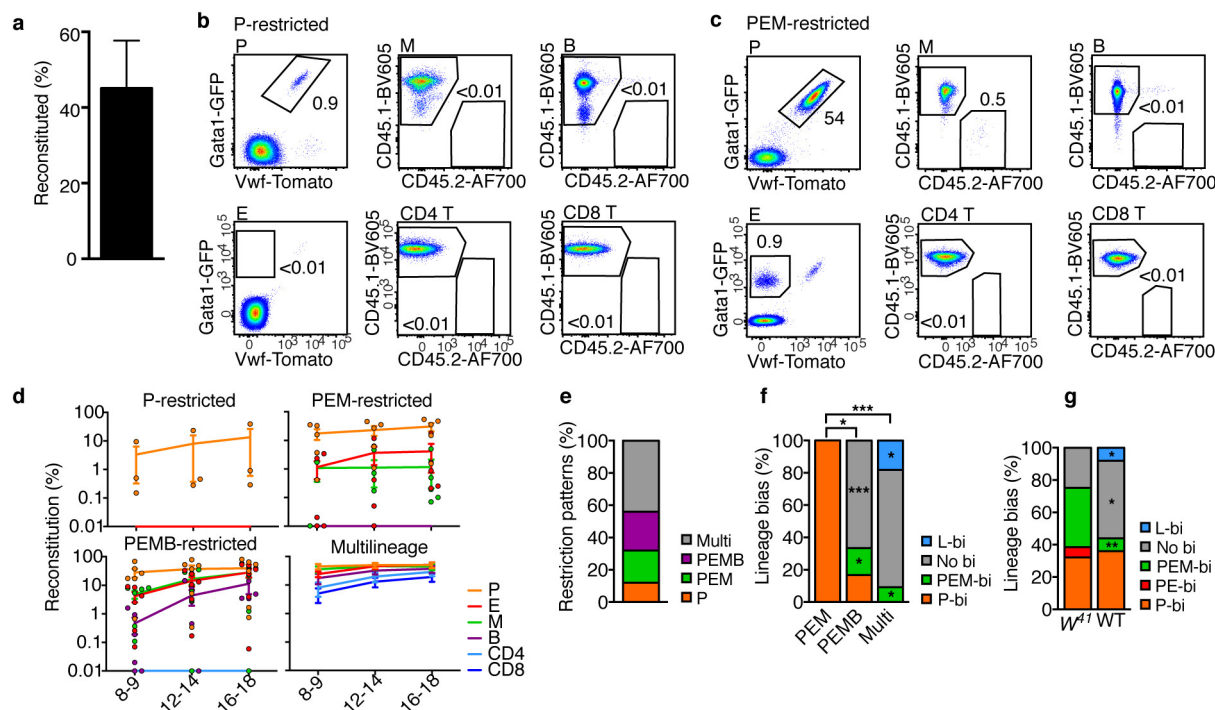
Data availability. Source data for all figures is available in the online version of the paper. Further details are available from the corresponding authors upon reasonable request.

31. Benveniste, P. *et al.* Intermediate-term hematopoietic stem cells with extended but time-limited reconstitution potential. *Cell Stem Cell* **6**, 48–58 (2010).
32. Tehranchi, R. *et al.* Persistent malignant stem cells in del(5q) myelodysplasia in remission. *N Engl J Med* **363**, 1025–1037 (2010).
33. Luis, T. C. *et al.* Initial seeding of the embryonic thymus by immune-restricted lympho-myeloid progenitors. *Nat Immunol* **17**, 1424–1435 (2016).
34. Marx, A., Backes, C., Meese, E., Lenhof, H. P. & Keller, A. EDISON-WMW: Exact dynamic programming solution of the Wilcoxon-Mann-Whitney test. *Genomics Proteomics Bioinformatics* **14**, 55–61 (2016).



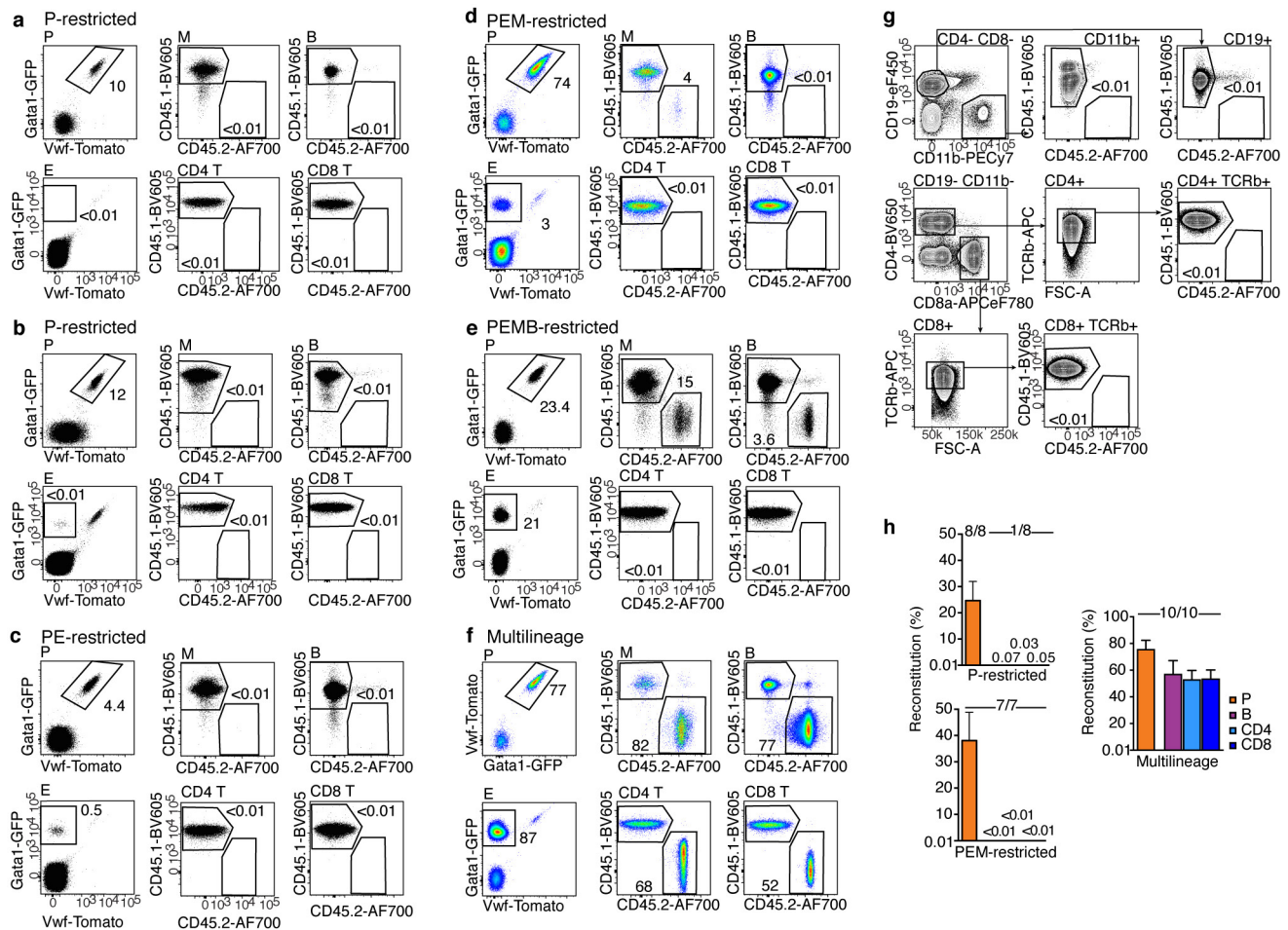
Extended Data Figure 1 | Characterization of *Vwf*-Tomato and *Vwf*-GFP co-expression in PB and BM. a-c, Flow cytometry analysis (1 experiment) of *Vwf*-Tomato and *Vwf*-GFP co-expression in PB platelets and erythroid cells (a), in c-Kit enriched BM LSK34⁺150⁺48⁻ cells showing representative

gating strategy used in single-cell transplantation sorts (b), and in myeloid and erythroid progenitors (c). d, FACS profile of BM LSK34⁺150⁺48⁻ cells in *Vwf*-tdTomato/*Gata1*-EGFP mice (2 experiments). Mean percentages of parent gates, n = 3 mice for all plots.



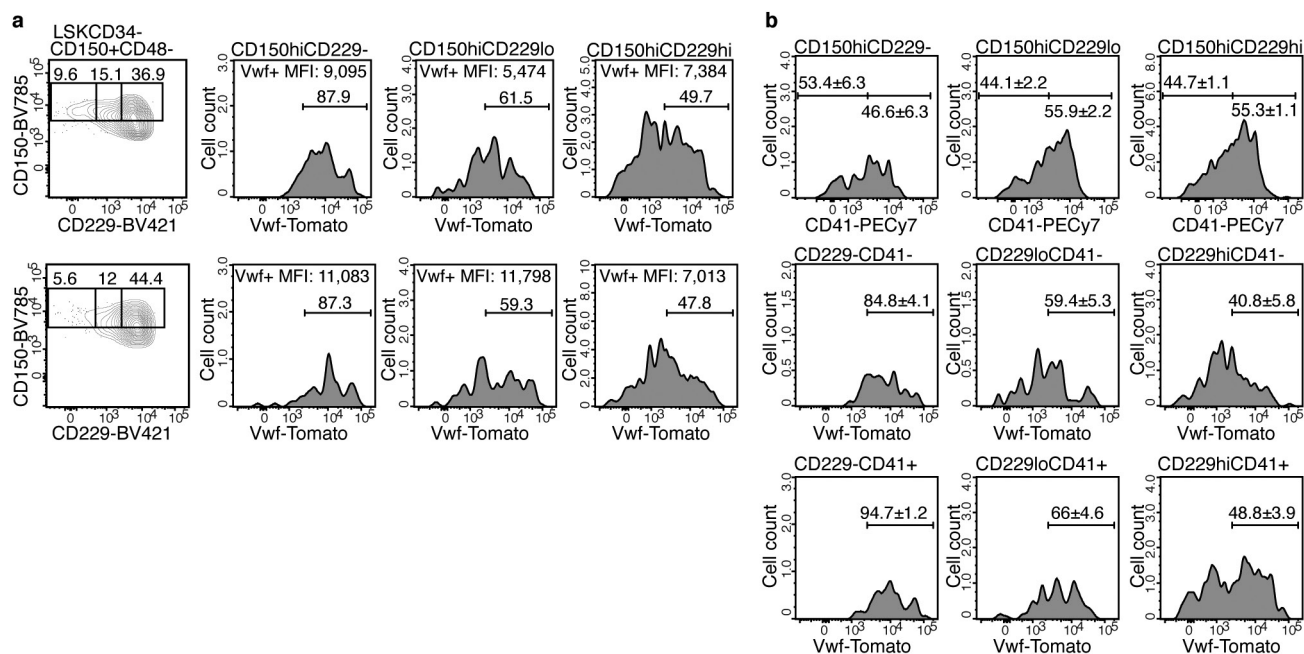
Extended Data Figure 2 | Stable long-term lineage-restricted reconstitution in recipients of single *Vwf-Tomato*^{pos} LSK34⁻150⁺48⁻ cells with wild-type BM support cells. **a**, Reconstituted mice (mean ± s.e.m.) transplanted with single *Vwf-Tomato*^{pos} LSK34⁻150⁺48⁻ cell and WT BM support. *n* = 58 transplanted mice, 7 experiments. No statistically significant difference in frequency of reconstituted mice between *W*⁴¹/*W*⁴¹ (Fig. 1a) and WT support (*p* = 0.46). **b–c**, Analysis of platelet, erythroid-, myeloid-, B- and T-cell contribution 16–18wks post-transplantation in P-restricted (**b**) and PEM-restricted mice (**c**) (plots representative of >40 single cell transplantation experiments). **d**, Reconstitution kinetics (mean ± s.e.m.). P-restricted, *n* = 3 mice; PEM-restricted, *n* = 5; PEMB-restricted, *n* = 6; Multilineage, *n* = 11. **e**, Distribution of lineage-restricted reconstitution patterns in mice in **a**. No statistically significant difference

in frequency of each pattern between mice transplanted with *W*⁴¹/*W*⁴¹ (Fig. 1d, *n* = 109 mice) and WT support (*n* = 25). Multi *p* = 0.662; PEMB *p* = 0.785; PEM *p* = 0.769; PE *p* = 1.0; P-restricted *p* = 1.0. **f**, Distribution of lineage-bias within reconstitution patterns in **e**. Statistically significant differences in P-bi frequency between patterns are indicated above bars (**p* = 0.015; ****p* = 0.0002); and for each pattern between mice transplanted with *W*⁴¹/*W*⁴¹ (Fig. 1e) and WT support within bars (PEMB: **p* = 0.022, ****p* = 0.0007; Multi: L-bi **p* = 0.026, PEM-bi **p* = 0.044). **g**, Overall lineage-bias distribution in **e** (WT support, *n* = 25 mice) and Fig. 1e (*W*⁴¹/*W*⁴¹ support, *n* = 109). L-bi **p* = 0.034; No bias **p* = 0.028; PEM-bi ***p* = 0.004. Statistical comparisons: two-tailed Fisher's exact test (95% CI).



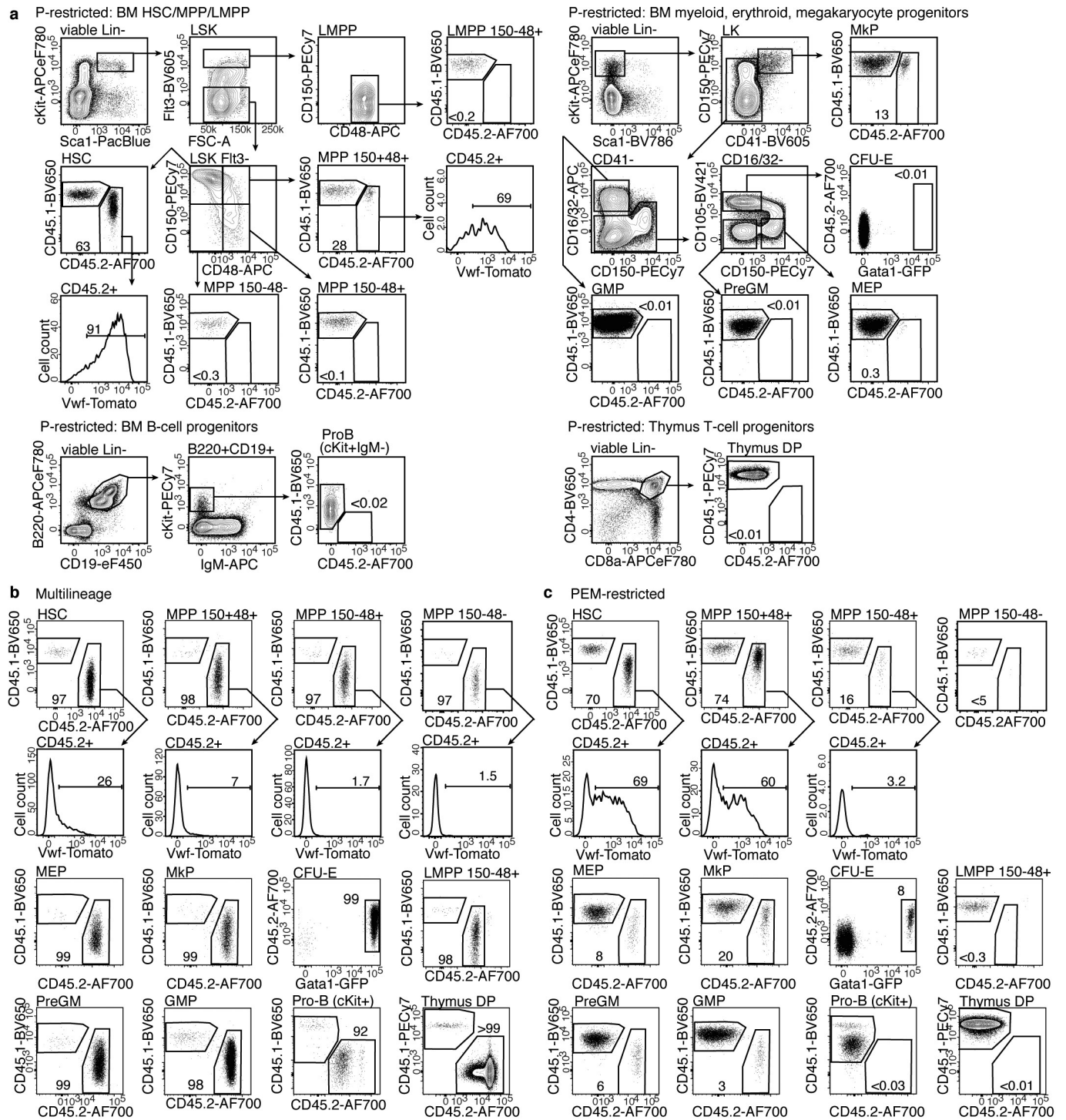
Extended Data Figure 3 | Analysis of PB lineage-restricted reconstitution by single *Vwf-Tomato*^{pos} LSK34⁺150⁺48⁺ cells and spleen lymphocyte reconstitution analysis in lineage-restricted reconstituted mice. **a-d**, Flow cytometry analysis of PB platelet, erythroid, myeloid, B and T lymphocyte reconstitution, 16–18wks post-transplantation of a single *Vwf-Tomato*^{pos} LSK34⁺150⁺48⁺ cell (representative of >40 single cell transplantation experiments). P-restricted (**a-b**), PE-restricted (**c**), PEM-restricted (**d**), PEMB-restricted (**e**) and Multilineage (**f**) stably

reconstituted mice. **g**, Flow cytometry analysis of myeloid and lymphoid reconstitution in spleen 23wks post-transplantation, corresponding to the P-restricted PB reconstitution pattern in **Extended Data Fig. 2b**. **h**, Reconstitution (mean±s.e.m.) of PB platelets and spleen lymphocytes 16–44wks post-transplantation in mice with P-restricted (n = 8 mice), PEM-restricted (n = 7) and Multilineage (n = 10) reconstitution, 13 experiments. Frequency of positive mice and mean reconstitution in positive mice shown.



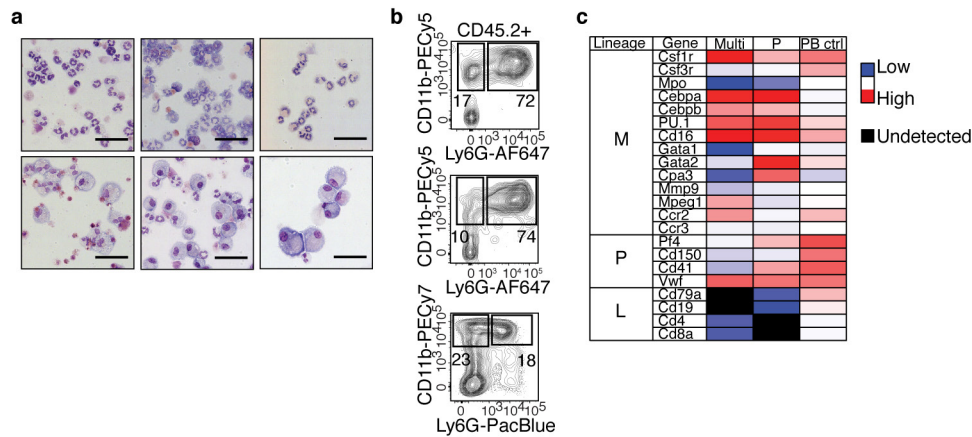
Extended Data Figure 4 | Vwf-Tomato co-expression with CD229 and CD41 in BM HSCs. a, Vwf-Tomato expression in LSK34⁻150^{hi}48⁻ cells with different CD229 expression levels in two different mice, representative of 4 experiments. Percentages of parent gate and median

fluorescence intensities (MFI) shown. **b,** Expression of CD41 and Vwf-Tomato (mean % of parent gate ± s.e.m., n = 3 mice in 1 experiment) in LSK34⁻150^{hi}48⁻ cells with different levels of CD229 expression.



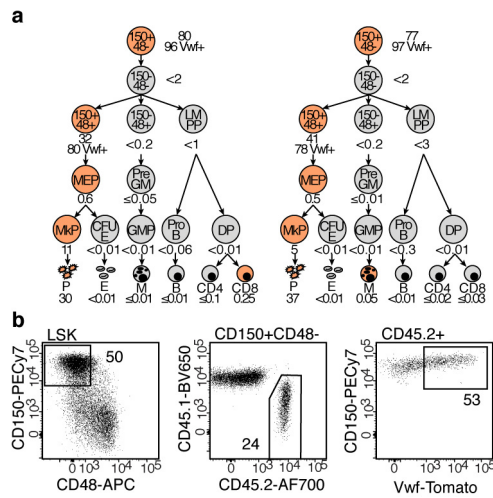
Extended Data Figure 5 | HSPC reconstitution analysis in lineage-restricted mice transplanted with single *Vwf-Tomato*^{POS} LSK34⁺ 150⁺48⁺ cells. a-d, Flow cytometry reconstitution analysis of BM and thymus HSPCs in a representative P-restricted reconstituted mouse at

23wks post-transplantation (a), and representative Multilineage (b) and PEM-restricted (c) reconstituted mice 16wks post-transplantation. Plots representative of 12 HSPC analysis experiments.



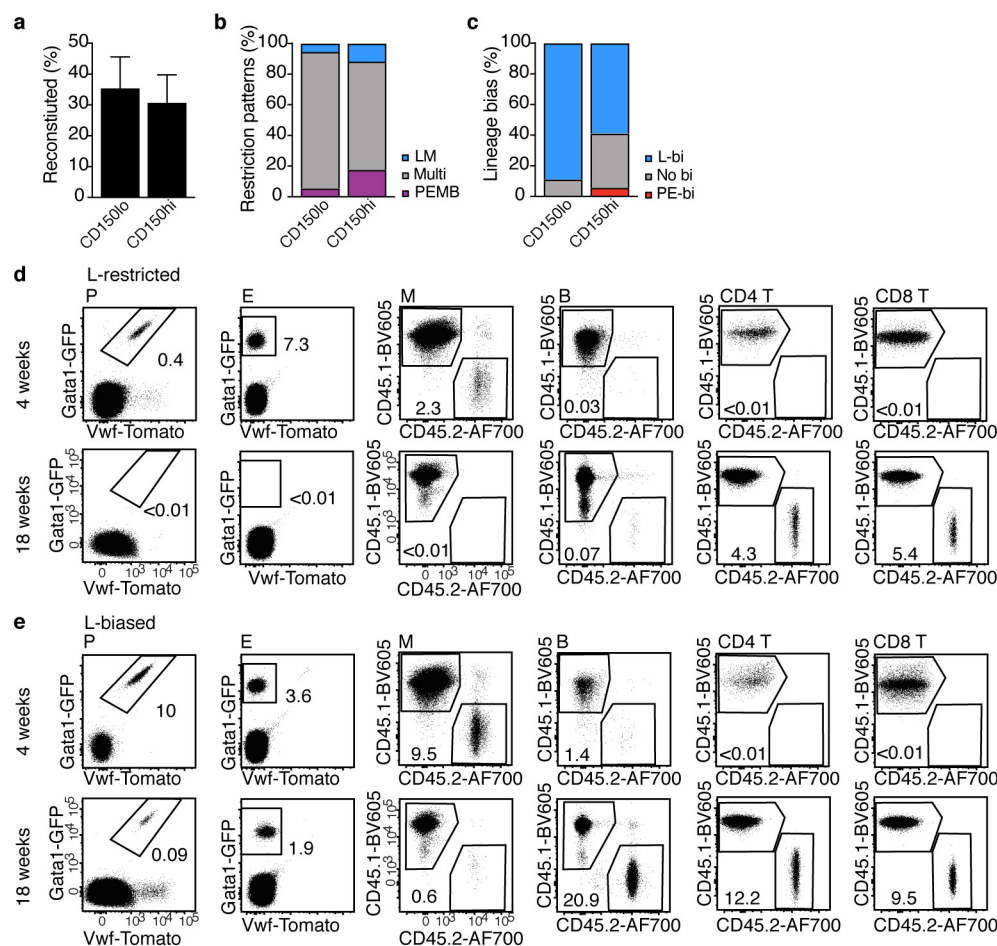
Extended Data Figure 6 | Multipotency of single reconstituting LSK34⁺150⁺48⁻ cells with *in vivo* lineage-restricted output. a, *In vitro*-derived granulocytes (top row) and monocytes/macrophages (bottom row) generated by donor-derived LSK cells sorted from single-HSC transplanted mice with long-term P-restricted reconstitution (each column shows results for one mouse; cytopins representative of 9 P-restricted mice analysed). Scale-bar = 50 μ m. b, FACS of granulocytes

(CD11b⁺Ly6G⁺) and monocytes/macrophages (CD11b⁺Ly6G⁻) generated by 25-50 donor-derived LSK from 3 P-restricted mice (representative of 4 mice analysed in 3 experiments). c, Gene expression heatmap of GM cells generated *in vitro*. Multi, n = 1 mouse. P-restricted, n = 2 mice. 1 experiment, 4 GM wells/mouse. Control: ~1,000 fresh PB leukocytes, n = 1. Mean Ct values per mouse group, normalized to mean Ct of *Hprt1/B2m*.



Extended Data Figure 7 | Long-term persistence of platelet-restricted and platelet-biased reconstitution patterns by multipotent HSCs.

a, Reconstitution of HSPC hierarchy in BM and thymus, 26wks post-transplantation, of two secondary recipients of a mouse with sustained P-restricted reconstitution at 44wks post-transplantation (additional recipient of the donor in **Fig. 3d**). **b**, Sorting strategy for secondary transplantation of *Vwf-Tomato*^{pos} HSCs from the mice in **Figure 3f** (representative gating of mouse 2). Percentages of parent gates shown.



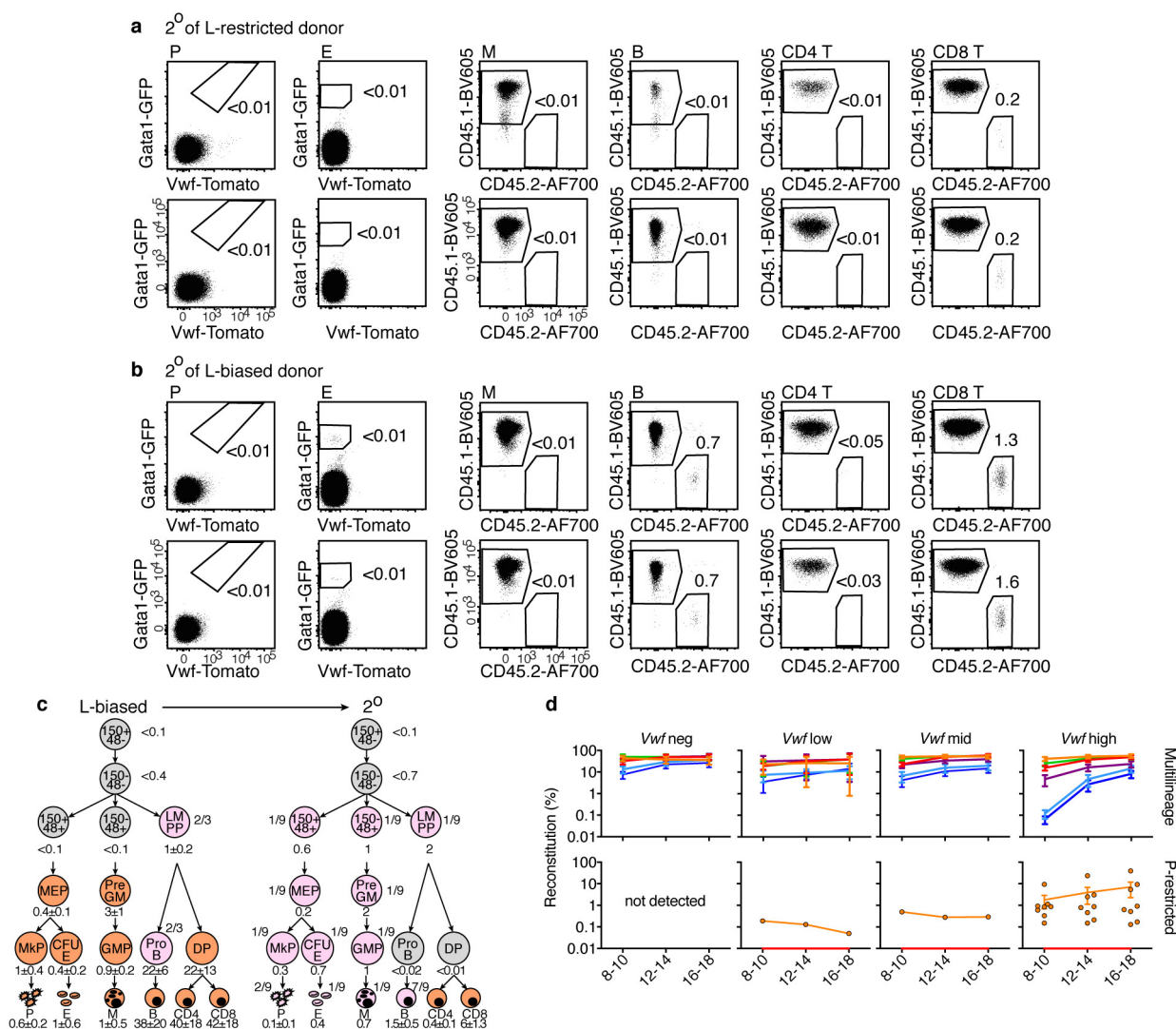
Extended Data Figure 8 | Reconstitution of PB mature lineages in mice transplanted with single LSK34⁺150⁺48⁻ cells with different levels of Vwf-Tomato expression and CD150 levels. **a**, Reconstituted mice (mean±s.e.m.) transplanted with single Vwf^{neg} HSCs with low (n=57) or high CD150 levels (n=77) in 11 experiments. No statistically significant difference, p=0.24. LSK34⁺150⁺48⁻ cells were index sorted and CD150^{hi}Vwf^{neg} cells defined as having CD150 expression levels overlapping with CD150 expression in LSK34⁺150⁺48⁻ Vwf^{mid-high} cells (see Fig.4a), and CD150^{lo}Vwf^{neg} cells defined as having lower CD150 expression levels than LSK34⁺150⁺48⁻ Vwf^{mid-high} cells. **b**, Distribution

of restriction patterns generated by single Vwf^{neg} CD150^{lo} (n=18 mice) and CD150^{hi} cells (n=17). No statistically significant differences. LM: p=0.60; Multi p=0.40; PEMB p=0.60. **c**, Distribution of lineage-bias in **e**. No statistically significant differences. No bias p=0.23; PE-bi p=0.49; L-bi p=0.12. **d-e**, Flow cytometry analysis of PB platelet, erythroid-, myeloid-, B- and T-cell reconstitution 4wks (top panels) and 18wks post-transplantation (bottom panels) in L-restricted (**b**) and L-biased (**c**) reconstituted mice (representative of 14 single cell transplantation experiments of Vwf^{neg} HSCs). Statistical comparisons: two-tailed Fisher's exact test (95% CI).



Extended Data Figure 9 | HSPC reconstitution analysis in L-restricted and L-biased reconstituted mice transplanted with a single LSK34⁺ 150⁺48⁻ cell. a-b, Flow cytometry analysis of PB platelet, erythroid-, myeloid-, B- and T-cell reconstitution, and HSPC hierarchy reconstitution, in L-restricted (a, representative of 3 experiments) and L-biased (b, representative of 2 experiments) reconstituted mice at 23wks

post-transplantation. c, Reconstitution percentages (mean ± s.e.m.) of HSPCs 22-23wks post-transplantation in L-restricted (n = 6 mice; except Pro-B, PreGM, GMP, MEP, CFU-E and MkP, n = 3) and L-biased (n = 3) reconstitution patterns. Orange: all mice positive. Grey: no positive mice. Pink: positive/negative mice; frequency and mean ± s.e.m. of positives shown.



Extended Data Figure 10 | Reconstitution analysis of secondary recipients of BM cells from single-cell transplanted mice with L-restricted and L-biased reconstitution. **a-b**, Flow cytometry analysis of PB platelet, erythroid-, myeloid-, B- and T-cell reconstitution 8wks post-transplantation of secondary recipients of BM cells from primary mice with L-restricted (**a**, representative of 3 experiments) and L-biased (**b**, representative of 2 experiments) reconstitution generated by a single LSK34⁺150⁺48⁻ cell. PB and BM HSPC reconstitution analysis of the primary recipients is shown in **Extended Data Fig. 9a-b**. **c**, Reconstitution

percentages (mean \pm s.e.m) of HSPCs in L-biased primary mice 21-29wks post-transplantation ($n = 3$) and their secondary recipients 17wks post-transplantation ($n = 9$, 3/donor). Orange: all mice positive. Grey: no positive mice. Pink: positive/negative mice; frequency and mean \pm s.e.m. of positives shown. **d**, Reconstitution kinetics (mean \pm s.e.m.) of mice in **Fig. 4h**, 14 experiments. Multilineage: Vwf^{neg} , $n = 6$ mice; Vwf^{low} , $n = 2$; Vwf^{mid} , $n = 12$; Vwf^{high} , $n = 12$. P-restricted: Vwf^{low} , $n = 1$; Vwf^{mid} , $n = 1$; Vwf^{high} , $n = 8$.

Tissue-selective effects of nucleolar stress and rDNA damage in developmental disorders

Eliezer Calo^{1,2}, Bo Gu³, Margot E. Bowen⁴, Fardin Aryan¹, Antoine Zalc³, Jialiang Liang¹, Ryan A. Flynn⁵, Tomek Swigut³, Howard Y. Chang⁶, Laura D. Attardi^{4,7} & Joanna Wysocka^{3,8,9}

Many craniofacial disorders are caused by heterozygous mutations in general regulators of housekeeping cellular functions such as transcription or ribosome biogenesis^{1,2}. Although it is understood that many of these malformations are a consequence of defects in cranial neural crest cells, a cell type that gives rise to most of the facial structures during embryogenesis^{3,4}, the mechanism underlying cell-type selectivity of these defects remains largely unknown. By exploring molecular functions of DDX21, a DEAD-box RNA helicase involved in control of both RNA polymerase (Pol) I- and II-dependent transcriptional arms of ribosome biogenesis⁵, we uncovered a previously unappreciated mechanism linking nucleolar dysfunction, ribosomal DNA (rDNA) damage, and craniofacial malformations. Here we demonstrate that genetic perturbations associated with Treacher Collins syndrome, a craniofacial disorder caused by heterozygous mutations in components of the Pol I transcriptional machinery or its cofactor TCOF1 (ref. 1), lead to relocalization of DDX21 from the nucleolus to the nucleoplasm, its loss from the chromatin targets, as well as inhibition of rRNA processing and downregulation of ribosomal protein gene transcription. These effects are cell-type-selective, cell-autonomous, and involve activation of p53 tumour-suppressor protein. We further show that cranial neural crest cells are sensitized to p53-mediated apoptosis, but blocking DDX21 loss from the nucleolus and chromatin rescues both the susceptibility to apoptosis and the craniofacial phenotypes associated with Treacher Collins syndrome. This mechanism is not restricted to cranial neural crest cells, as blood formation is also hypersensitive to loss of DDX21 functions. Accordingly, ribosomal gene perturbations associated with Diamond-Blackfan anaemia disrupt DDX21 localization. At the molecular level, we demonstrate that impaired rRNA synthesis elicits a DNA damage response, and that rDNA damage results in tissue-selective and dosage-dependent effects on craniofacial development. Taken together, our findings illustrate how disruption in general regulators that compromise nucleolar homeostasis can result in tissue-selective malformations.

Heterozygous mutations in factors involved in ribosome biogenesis lead to ribosomopathies⁶, a collection of congenital disorders typically displaying tissue-selective defects, despite the broad requirement for ribosomes across growing tissues. For example, Treacher Collins syndrome (TCS), caused by heterozygous mutations in Pol I cofactor TCOF1 or subunits POLR1D and POLR1C, is characterized by a specific set of craniofacial malformations⁷. To explore the mechanism by which perturbations in ribosomal gene transcription result in TCS, we focused on DDX21, a nucleolar protein involved in the control of the two transcriptional arms of ribosome biogenesis: (1) synthesis and processing of the rRNA in the nucleolus, and (2) transcription of

ribosomal protein genes in the nucleoplasm⁵. Induction of nucleolar stress by inhibition of Pol I leads to DDX21 relocalization from the nucleolus to the nucleoplasm and to its simultaneous loss from Pol I and Pol II target promoters⁵. Furthermore, single-cell measurements revealed a strong correlation between the DDX21 nucleolar/nucleoplasmic ratio and pre-rRNA levels, both in unperturbed HeLa cells and in those treated with the Pol I inhibitor CX-5461 (hereafter iPol I) (Fig. 1a, b).

We asked whether perturbations in TCS-associated genes elicit disruption of DDX21 functions. Downregulation of *TCOF1* or *POLR1D* in HeLa cells (Extended Data Fig. 1b, c) led to relocalization of DDX21 to the nucleoplasm (Fig. 1c and Extended Data Fig. 1d), and this was accompanied by eviction of DDX21 from the rDNA and Pol II target promoters, as determined by chromatin immunoprecipitation followed by sequencing (ChIP-seq) (Fig. 1d–f) and confirmed in independent ChIP-qPCR experiments (Extended Data Fig. 1e, f). Analysis of *TCOF1* genomic occupancy showed that although it binds the rDNA (Fig. 1g), unlike DDX21, it does not associate with Pol II promoters (Fig. 1h, i). Even within the nucleolus, DDX21 and *TCOF1* may not act as a part of the same complex, as they do not readily co-immunoprecipitate (Extended Data Fig. 2a, b). Taken together, our data suggest that DDX21 can respond to *TCOF1* dysfunction indirectly, through a pathway that is sensitive to the status of rRNA synthesis.

TCS craniofacial anomalies originate primarily from diminished allocation of cranial neural crest cells (cNCCs) into the first and second pharyngeal arches^{3,4,7}. If loss of DDX21 from chromatin is an important downstream mediator of *TCOF1* and *POLR1D* dysfunction, then first and second arch structures should be sensitive to DDX21 knockdown. To test this, we established *Xenopus* as a model for TCS. *tcof1* knockdown in *Xenopus* embryos with morpholinos targeting either translation or splicing of *tcof1* resulted in hypoplasia and deformation of the mandibular and hyoid stream cartilage structures, which are derived from first and second arches (Fig. 1j and Extended Data Fig. 2c–e), a phenotype consistent with both TCS and published zebrafish phenotypes^{8–10}. Notably, at higher morpholino doses, overall growth defects were evident in *tcof1* morphants (Extended Data Fig. 2e). Next, we designed and injected morpholino targeting *ddx21*. At high doses, this morpholino impeded embryonic growth, but at lower doses we observed craniofacial phenotypes remarkably similar to those seen in *tcof1* morphants (Fig. 1j and Extended Data Fig. 2c, e). Furthermore, injection of mRNA encoding a catalytically defective human DDX21 (ref. 11) (DDX21^{SAT}) also faithfully phenocopied TCS craniofacial defects, whereas injection of wild-type DDX21 mRNA had no appreciable effects (Fig. 1j and Extended Data Fig. 2c, e, f). Thus, development of craniofacial structures is hypersensitive to the loss of Ddx21 or its RNA helicase activity. These results were corroborated in

¹Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²David H. Koch Institute for Integrative Cancer Research, Cambridge, Massachusetts 02139, USA. ³Department of Chemical and Systems Biology, Stanford University School of Medicine, Stanford, California 94305, USA. ⁴Department of Radiation Oncology, Division of Radiation and Cancer Biology, Stanford University School of Medicine, Stanford, California 94305, USA. ⁵Department of Chemistry, Stanford University, Stanford, California 94305, USA. ⁶Center for Personal Dynamic Regulomes, Stanford University, 269 Campus Drive, Stanford, California 94305, USA. ⁷Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. ⁸Department of Developmental Biology, Stanford University School of Medicine, Stanford, California 94305, USA. ⁹Howard Hughes Medical Institute, Stanford School of Medicine, Stanford University, Stanford, California 94305, USA.

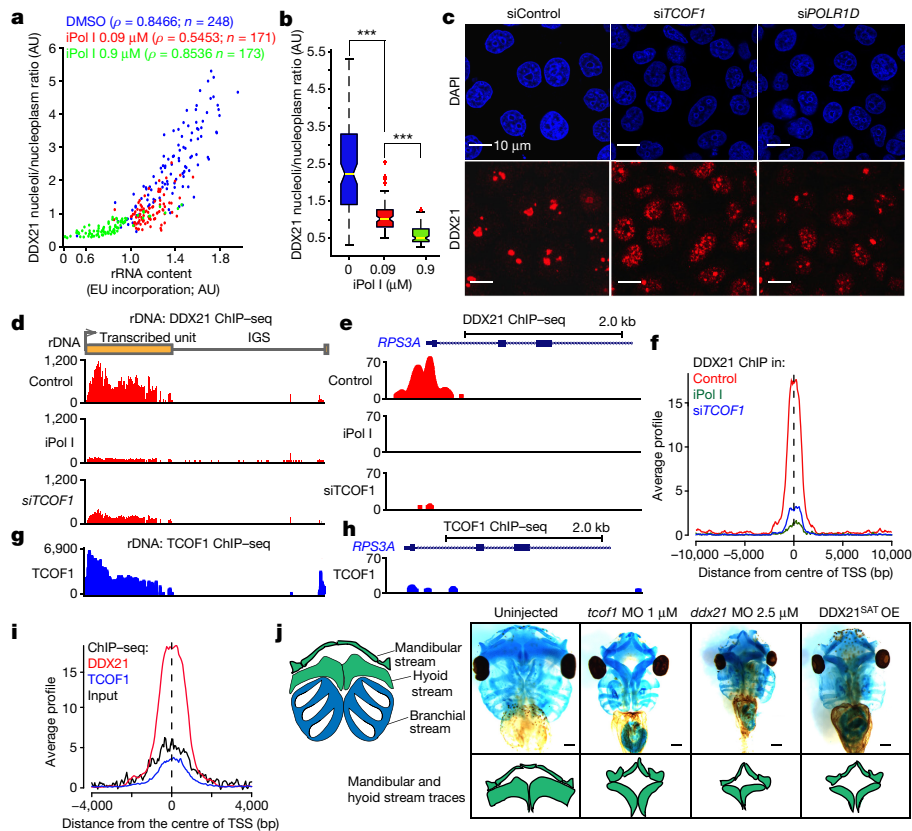


Figure 1 | The functions of DDX21 are linked to rRNA synthesis levels and altered by TCS-associated perturbations. **a, b,** Quantification of the relationship between DDX21 nucleolar/nucleoplasmic ratio and/or pre-rRNA synthesis after 1 h treatment of HeLa cells with different dosages of iPol I. Cells were collected from $n = 3$ biologically independent experiments. ρ , Pearson correlation coefficient; EU, 5-ethynyl uridine; AU, arbitrary units. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. $***P < 0.001$, two-sided Wilcoxon–Mann–Whitney test. **c,** Representative immunofluorescence images depicting DDX21 localization changes upon siRNA-mediated knockdown of *TCOF1* (siTCOF1) or *POLR1D* (siPOLR1D). $n = 3$ biologically independent experiments. **d,** Mapping of DDX21 ChIP-seq reads, from HeLa cells treated with dimethylsulfoxide

(DMSO), iPol I, or siTCOF1, to the rDNA locus. **e,** Genome browser tracks depicting DDX21 ChIP-seq signal, from HeLa cells treated with DMSO, iPol I, or siTCOF1, at the *RPS3A* locus. **f,** Average signal profiles of DDX21 ChIP-seq from cells treated with DMSO, iPol I, or siTCOF1. **g,** TCOF1 ChIP-seq reads mapped to the rDNA locus. **h,** Genome browser tracks of TCOF1 ChIP-seq signal at the *RPS3A* locus. **i,** Average signal profiles comparing DDX21 (same as in **f**) and TCOF1 ChIP-seq, and background input reads. ChIP-seq has been extensively validated by ChIP-qPCR and in another cell type (data not shown and ref. 5). **j,** Representative stainings of *Xenopus laevis* cranial cartilages at stage 49. Traces display the mandibular and hyoid stream defects. MO, morpholino; OE, overexpression. Animals were collected from $n = 3$ biologically independent experiments.

zebrafish embryos injected with morpholinos blocking translation of *ddx21*, where observed craniofacial phenotypes resembled those reported in the *polr1d*^{−/−} and *polr1c*^{−/−} models of TCS^{8–10}, and were rescued by the co-injection of human *DDX21* mRNA (Extended Data Fig. 2g–j).

Because TCS phenotypes were shown to result from defects in cNCCs^{9,10,12,13}, we investigated whether *Tcof1* loss cell-autonomously affects DDX21 functions in cNCCs. We used CRISPR–Cas9 genome editing to generate *Tcof1*^{+/−} and *Tcof1*^{−/−} mouse embryonic stem (ES) cells, which showed no appreciable defect in Ddx21 nucleolar localization (Fig. 2a, left, and Extended Data Fig. 3a). However, upon differentiation into cNCCs, we observed partial relocalization of Ddx21 to the nucleoplasm in *Tcof1*-mutant, but not in wild-type cells (Fig. 2a, right). This defect was rescued by introduction of an inducible human GFP-tagged *TCOF1* construct (GFP–TCOF1) (Fig. 2b and Extended Data Fig. 3b). The severity of the defect was dependent on *Tcof1* dosage, with more pronounced nucleolar exclusion observed in *Tcof1*^{−/−} cNCCs (Fig. 2a). Ddx21 relocalization was not observed in embryoid body outgrowths from *Tcof1*^{−/−} (Extended Data Fig. 3c), further suggesting cell-type selectivity. We also generated a *TCOF1*^{+/−} human ES cell line (Extended Data Fig. 3d–f) and observed partial relocalization of DDX21 only after differentiation of *TCOF1*^{+/−} human ES cells to cNCCs (Extended Data Fig. 3g).

Consistent with cNCC-selective DDX21 relocalization, we also observed partial loss of DDX21 from rDNA and its Pol II target promoters in *Tcof1*^{−/−} cNCCs, but not in mouse ES cells (Fig. 2c, d). Similar depletion of DDX21 from chromatin occurred in human *TCOF1*^{+/−} cNCCs (Extended Data Fig. 3h) and was accompanied by down-regulation of DDX21-bound Pol I and Pol II target genes (Extended Data Fig. 3i). In addition to transcription, DDX21 is also required for rRNA processing^{5,14} through its interaction with both rRNA and small nucleolar RNAs (snoRNAs). Pol I inhibition disengaged DDX21 from both the 5' external transcribed spacer, a site of processing in the rRNA, and from the snoRNAs (Extended Data Fig. 4a–e). Accordingly, *Tcof1*-mutant cells also display cell-type-selective impairment of the 5' external transcribed spacer processing, specifically cleavage of the A' site, which is one of the first events during maturation of the 18S rRNA and only accumulates when rRNA processing is defective (Fig. 2e). Accumulation of unprocessed A' site was both dosage-dependent and rescued by GFP–TCOF1 overexpression (Fig. 2e). Taken together, our results demonstrate that TCOF1 mutations result in cell-type-selective and cell-autonomous perturbations of DDX21 functions in cNCCs.

In addition to rRNA synthesis and processing defects¹⁵, TCS is also characterized by nucleolar stress-mediated activation of p53 (refs 10, 16). Accordingly, we observed upregulation of the canonical p53 target *Cdkn1a* (*p21*) in *Tcof1*-deficient mouse and

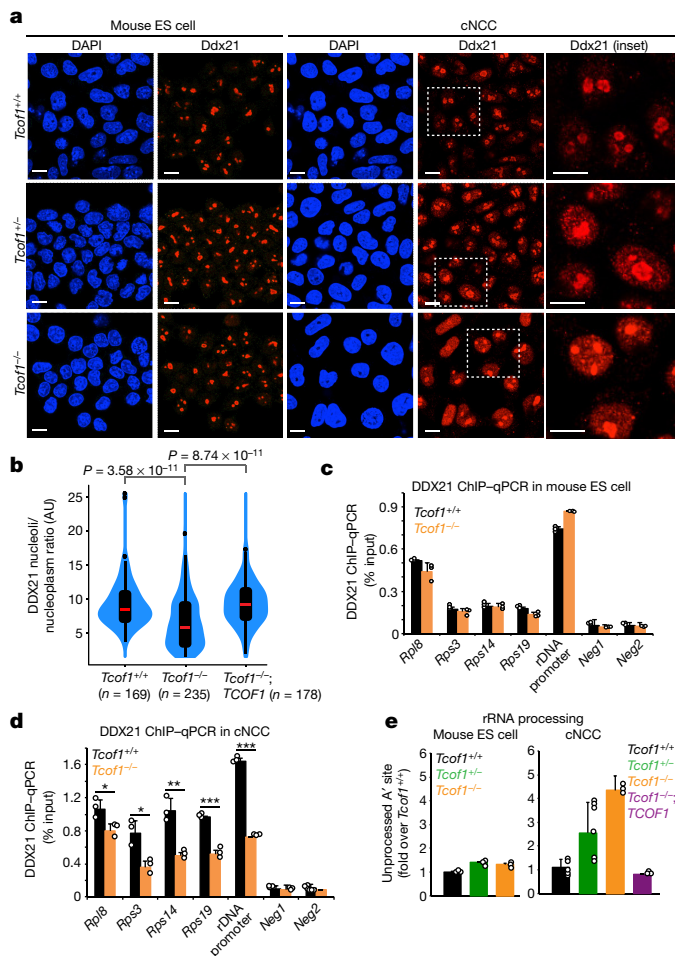


Figure 2 | DDX21 deregulation in TCS is both cell-autonomous and cell-type selective. **a**, Representative immunofluorescence images of DAPI and DDX21 in mouse ES cell and mouse ES cell-derived cNCCs of wild-type, *Tcof1*^{+/+}, or *Tcof1*^{-/-} genotypes. *n* = 3 biologically independent experiments. Scale bars, 10 μm. **b**, Violin plots quantifying DDX21 nucleolar/nucleoplasmic ratio in cNCCs of wild-type, *Tcof1*^{+/-}, or *Tcof1*^{-/-} rescued by expression of an inducible human GFP-TCOF1 construct. Cells were collected from three biologically independent experiments. Boxes represent median value (red line) and 25th and 75th percentiles, whiskers are minimum to maximum, points are outliers. Two-sided Wilcoxon–Mann–Whitney test. **c**, **d**, ChIP–qPCR analysis from mouse ES cells and cNCCs sampling Ddx21 genomic occupancy, at a representative panel of Ddx21 targets. **e**, qPCR analyses examining processing of the 47S rRNA A' site in mouse ES cells and cNCCs. **c–e**, Bars represent the average of *n* = 3 biologically independent experiments; error bars, s.e.m. ****P* < 0.001, ***P* < 0.01, **P* < 0.05, two-sided, unpaired *t*-test.

TCOF1-deficient human cNCCs (Extended Data Fig. 5a, b). Treatment of wild-type cNCCs with NSC146109, a small molecule that promotes p53 stabilization^{17,18} (Extended Data Fig. 5c), resulted in DDX21 relocalization (Fig. 3a, b) and concomitant inhibition of rRNA synthesis (Fig. 3c), suggesting that p53 activation is sufficient to induce DDX21 nucleolar exclusion. This is also true *in vivo*, as dorsal–anterior neural tubes of mouse embryos with the neural-crest-specific knockout of *Mdm2* (*Wnt1-cre;Mdm2*^{fl/fl}) showed relocalization of DDX21 only in those cells containing high levels of p53, corresponding to the developing cNCCs (Fig. 3d and Extended Data Fig. 5d).

Despite a broad function of p53 as a stress sensor, stabilizing *p53* mutations lead to specific craniofacial defects in mice¹⁹. We investigated whether *p53* mRNA levels are elevated in the developing cNCCs, which could result in a larger reservoir of the translated product being available for stabilization upon stress, when the E3 ligase-mediated

degradation of p53 is no longer the limiting regulatory step. Though *p53* mRNA is broadly expressed throughout the mouse embryo, substantially higher levels are detected in the neural tube and craniofacial region, especially the first arch (Fig. 3e and Extended Data Fig. 5e–g). Furthermore, elevated *p53* mRNA levels have previously been observed in the dorsal neural tubes of chick embryos before the onset of cNCC emigration²⁰. This elevated expression could potentially contribute to the sensitivity of cNCCs and craniofacial development to p53 activity upon stress, although other mechanisms may also be at play. Regardless, *Wnt1-cre;Mdm2*^{fl/fl} embryos showed hypoplasia of the first and second pharyngeal arches, confirming that these facial structures are sensitive to p53 stabilization (Extended Data Fig. 6).

To directly address whether human cNCCs are more sensitive to p53 stabilization than other embryonic cell types, we differentiated human ES cells into endothelial cells, cardiomyocytes, and cNCCs, and further differentiated cNCCs to smooth muscle cells. Quantification of apoptosis after parallel treatment of this isogenic set of cell types with NSC146109 revealed the highest sensitivity of cNCCs (Fig. 3f). This effect was not due to changes in DDX21 protein level among the different cell types (Extended Data Fig. 7a). Moreover, loss of one *TCOF1* allele exacerbated this sensitivity (Extended Data Fig. 7b).

We next investigated whether disruption of DDX21 underlies cNCC sensitivity to p53. Indeed, induction of DDX21 overexpression partly rescued cNCC sensitivity to p53 activation (Fig. 3g). Moreover, overexpression of DDX21 mRNA rescued the cranial cartilage defects associated with *TCOF1* dysfunction in *Xenopus* embryos to a similar extent as *TCOF1* mRNA or *p53* knockdown (Extended Data Fig. 7c, d). Consistently, overexpression of DDX21 in TCS cNCCs rescued DDX21 nucleolar localization and chromatin association (Extended Data Fig. 7e, f). Thus, preventing DDX21 loss from nucleolus and chromatin suppresses sensitivity of cNCCs to apoptosis and developmental defects associated with TCS.

Given that nucleolar stress and p53 activation are hallmarks of many ribosomopathies⁶, we reasoned that DDX21 dysfunction could contribute to these disorders. To explore this, we performed a small-scale short interfering RNA (siRNA) screen in HeLa cells to assess DDX21 nucleolar localization upon knockdown of a subset of ribosomopathy-associated genes (Fig. 3h and Extended Data Fig. 7h, g). We observed that knockdown of genes implicated in TCS, Diamond–Blackfan anaemia (*RPS19*) and Shwachman–Diamond syndrome (*SDS*) caused DDX21 relocalization, while others had no effect (Fig. 3h). Although patients with Diamond–Blackfan anaemia and Shwachman–Diamond syndrome develop craniofacial/skeletal deformities, these syndromes are characterized primarily by bone marrow dysfunction^{1,6}. To test whether loss of DDX21 results in anaemia, we performed haemoglobin stainings of *ddx21* zebrafish morphants. Similar to embryos derived from the *rpl11*^{-/-} zebrafish model of Diamond–Blackfan anaemia²¹, *ddx21* morphants were anaemic (Fig. 3i and Extended Data Fig. 7i), suggesting that hypersensitivities to loss of DDX21 exist beyond the craniofacial development and may contribute to the pathogenesis of multiple ribosomopathies.

TCOF1 dysfunction has also been linked to elevated DNA damage²². To explore potential relationships between Pol I transcriptional stress, DNA damage, and DDX21, we examined induction of γH2A.X in *TCOF1*^{+/+} cNCCs, in cNCCs treated with iPol I or low levels of actinomycin D, and in HeLa cells treated with iPol I. In all cases, we observed elevated γH2A.X signals in a subset of cells (Extended Data Fig. 8a–g).

DNA damage induction in response to Pol I inhibition also occurs *in vivo*, as we detected elevated γH2A.X signals in *polr1d*^{-/-} and *polr1c*^{-/-} zebrafish embryos at 24 h post-fertilization (Fig. 4a). The staining was more pronounced in the anterior parts of the embryo, but also evident in non-cranial tissues (Fig. 4a). However, the severity of cranial cartilage malformations was correlated with the γH2A.X signals (Fig. 4a–c).

We hypothesized that rDNA damage caused by Pol I transcriptional stress may be a trigger for p53 activation and DDX21 relocalization.

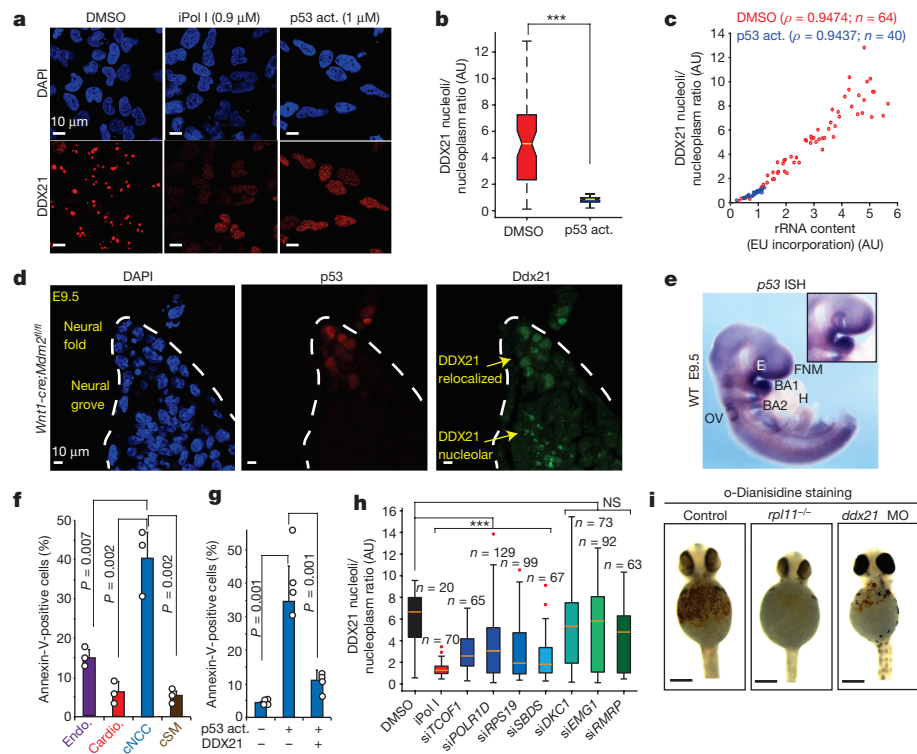


Figure 3 | Selective sensitivity of cNCCs to p53 activation (act.) and DDX21 levels. **a**, Representative immunofluorescence images of DAPI and DDX21 stainings in human cNCCs treated with DMSO, iPol I or the p53 activator (NSC146109). $n = 3$ biologically independent experiments. **b**, **c**, Quantification of DDX21 nucleolar/nucleoplasmic ratio (**b**) and the relationship between rRNA synthesis and DDX21 localization (**c**) in cNCCs after 12 h treatment with DMSO or NSC146109. Cells were collected from three biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum. *** $P < 0.001$, two-sided Wilcoxon–Mann–Whitney test. **d**, Immunofluorescence staining of p53 and DDX21 in sections from the dorsal neural tube of *Wnt1-cre;Mdm2^{fl/fl}* embryonic day (E)9.5 mouse embryos. Dotted lines represent the neural fold. $n = 5$ independent animals per genotype. **e**, Representative picture of whole-mount *in situ* hybridization (ISH) for p53 mRNA in E9.5 embryos. Inset highlights the first brachial arch (BA1). WT, wild type; E, eye; FNM, frontonasal mass; H, heart; OV, otic vesicle. $n = 4$ independent animals. **f**, Differentiation of human ES cells into

the indicated cell types (Endo., endothelial cells; Cardio., cardiomyocytes; cSM, cNCC-derived smooth muscle). Sensitivity to p53-mediated apoptosis after treatment with NSC146109 for 12–16 h was quantified by fluorescence-activated cell sorting (FACS). Bars are from $n = 3$ biologically independent experiments; error bars, s.e.m. **g**, Human cNCC wild type or overexpressing GFP-DDX21 were treated with NSC146109, followed by FACS analyses of annexin V staining. Bars are from $n = 4$ biologically independent experiments; error bars, s.e.m. For **f** and **g** P values, two-sided, unpaired *t*-test. **h**, Box plot represents quantification of DDX21 nucleolar/nucleoplasmic ratio for indicated siRNAs. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. *** $P < 0.001$, two-sided Wilcoxon–Mann–Whitney test. NS, not significant. Cells were collected from three biologically independent experiments. **i**, Haemoglobin staining with o-dianisidine of control, *rpl11^{-/-}*, and *ddx21* morpholino-injected zebrafish embryos at 24 h post-fertilization. Embryos were collected from $n = 3$ independent matings.

To test this, we used tamoxifen-inducible U2OS lines expressing either I-PpoI, an endonuclease which preferentially cleaves the rDNA^{23,24}, or AsiSI, which causes DNA damage outside rDNA²⁵. Consistently, the γ H2A.X signal was predominantly perinucleolar in I-PpoI-treated cells, and nucleoplasmic in AsiSI-treated cells (Fig. 4d). Only cells expressing I-PpoI relocalized DDX21 (Fig. 4d, e). Furthermore, although both I-PpoI and AsiSI induced auto-phosphorylation of DNA-dependent protein kinase catalytic subunit (pDNA-PKCs) and stabilized p53 (Fig. 4f), rDNA damage induced p53 protein to higher levels (Fig. 4f), despite overall higher γ H2A.X levels in AsiSI-treated cells (Extended Data Fig. 8k). Single-cell analysis revealed that p53 stabilization is much more correlated with the number of the DNA damage foci in I-PpoI than in AsiSI-treated cells, suggesting an intricate link between the rDNA damage and p53 stabilization (Extended Data Fig. 8h–j).

We next investigated temporal relationships and dependencies between Pol I inhibition, DNA damage signalling, and DDX21 relocalization. We observed a gradual increase in pDNA-PKCs within 15–30 min of iPol I treatment, coinciding with timing of DDX21 relocalization (Extended Data Fig. 9a–c). Inhibition of DNA-PKCs and ataxia–telangiectasia mutated (ATM) abolished the displacement of DDX21 from the nucleolus upon rDNA damage (Extended Data Fig. 9d, e). We also observed a reciprocal dependency, as knockdown

of DDX21 resulted in elevated pDNA-PKCs (Extended Data Fig. 9f), consistent with a cross-talk between DDX21, DNA damage, and Pol I transcription²⁶.

If rDNA damage contributes to TCS pathology, it would suggest that even a general insult such as rDNA damage can result in tissue-selective phenotypes. We leveraged that I-PpoI cleavage site is conserved in *Xenopus* rDNA and injected various amounts of *I-PpoI* mRNA into *Xenopus* embryos. Overall growth and morphology of the resulting tadpoles appeared relatively unaffected with exception of the head development (Extended Data Fig. 9g). However, analysis of the cranial cartilage showed that low doses of *I-PpoI* mRNA result in phenotypes resembling those seen in TCS and upon *ddx21* knock-down (Fig. 4g). At higher doses, we observed a more severe and highly variable spectrum of phenotypes that were predominantly limited to the head region, although an overall effect on growth was also evident (Fig. 4g and Extended Data Fig. 9h).

Because craniofacial development is selectively sensitive to rDNA damage, our results suggest that phenotypic variability observed in patients with TCS with the same mutation might be due, among other factors, to variable levels of rDNA damage resulting from *in utero* exposure, genetic modifiers, and/or stochastic events. Furthermore, various genetic and/or environmental perturbations leading to defects

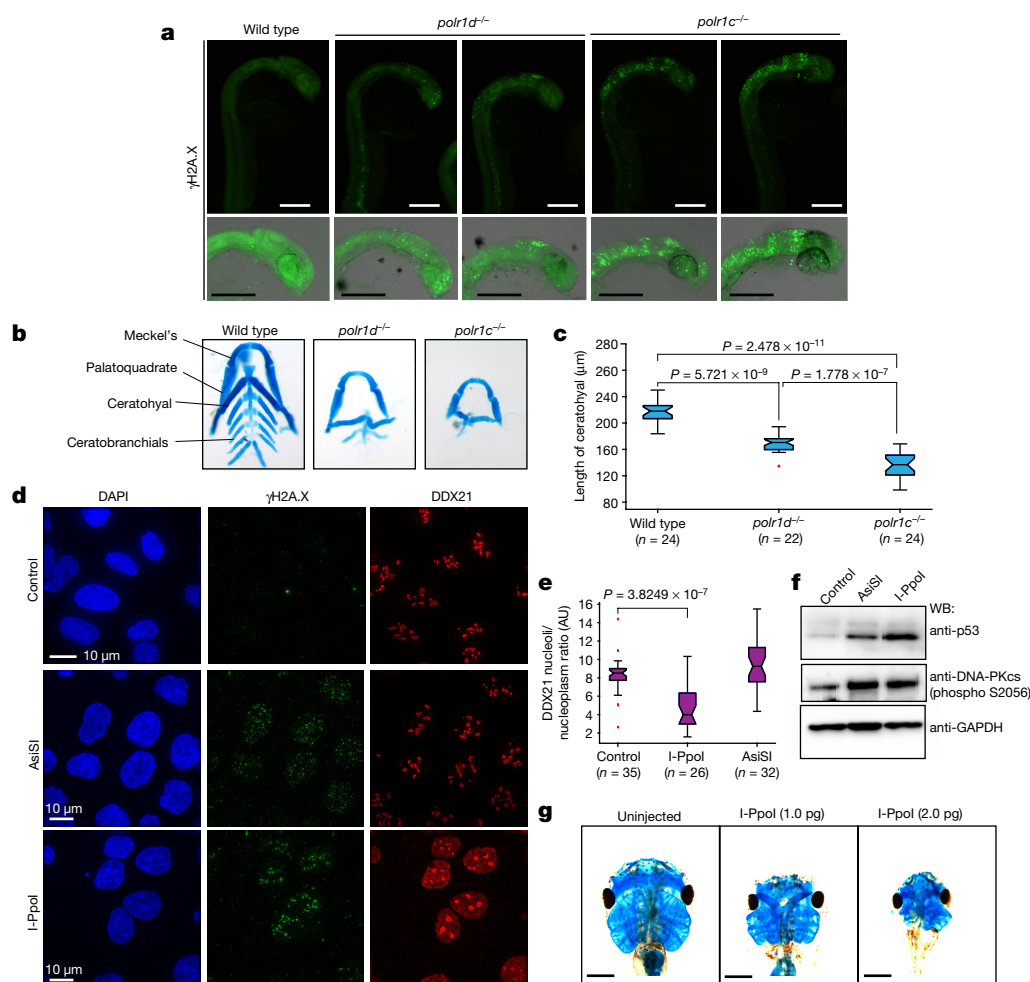


Figure 4 | rDNA damage induces DDX21 relocalization and impairs craniofacial development. **a**, Representative γ H2A.X staining of wild-type, *polr1d*^{-/-}, and *polr1c*^{-/-} zebrafish embryos. Embryos were stained from *n* = 3 independent matings. **b**, Representative images of dissected wild-type, *polr1d*^{-/-}, and *polr1c*^{-/-} zebrafish craniofacial cartilages. Embryos were collected and stained from *n* = 3 independent matings. **c**, Box plot depicting the length of the ceratohyal cartilage of wild-type, *polr1d*^{-/-}, and *polr1c*^{-/-} zebrafish embryos. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. *P* values, two-sided Wilcoxon–Mann–Whitney test. **d**, Representative immunofluorescence of U2OS

expressing or not AsiSI or I-PpoI from *n* = 4 biologically independent experiments. **e**, Box plot quantifying DDX21 nucleolar/nucleoplasmic ratio in cells expressing AsiSI or I-PpoI. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. *P* values, two-sided Wilcoxon–Mann–Whitney test. **f**, Representative western blots from cells expressing or not AsiSI or I-PpoI for 4 h from *n* = 3 independent biological experiments. **g**, Representative stainings of *Xenopus* cranial cartilages from embryos injected with different dosages of *I-PpoI* mRNA from *n* = 4 biologically independent experiments.

in rRNA synthesis and rDNA damage will probably be associated with craniofacial malformations (Extended Data Fig. 10). In this context, our results provide a unified molecular framework that has the potential to explain, at least in part, why craniofacial malformations are very common, often arise through mutations in generic regulators, and show high sensitivity to environmental modulation and phenotypic variability among affected individuals.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 November 2016; accepted 11 December 2017.

Published online 24 January 2018.

- Yelick, P. C. & Trainor, P. A. Ribosomopathies: global process, tissue specific defects. *Rare Dis.* **3**, e1025185 (2015).
- Berdasco, M. & Esteller, M. Genetic syndromes caused by mutations in epigenetic genes. *Hum. Genet.* **132**, 359–383 (2013).
- Trainor, P. A. Craniofacial birth defects: the role of neural crest cells in the etiology and pathogenesis of Treacher Collins syndrome and the potential for prevention. *Am. J. Med. Genet. A* **152**, 2984–2994 (2010).

- Bronner, M. E. & LeDouarin, N. M. Development and evolution of the neural crest: an overview. *Dev. Biol.* **366**, 2–9 (2012).
- Calo, E. *et al.* RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. *Nature* **518**, 249–253 (2015).
- Narla, A. & Ebert, B. L. Ribosomopathies: human disorders of ribosome dysfunction. *Blood* **115**, 3196–3205 (2010).
- Kadakia, S., Helman, S. N., Badhey, A. K., Saman, M. & Ducic, Y. Treacher Collins syndrome: the genetics of a craniofacial disease. *Int. J. Pediatr. Otorhinolaryngol.* **78**, 893–898 (2014).
- Weiner, A. M. J., Scamporrì, N. L. & Calcaterra, N. B. Fishing the molecular bases of Treacher Collins syndrome. *PLoS ONE* **7**, e29574 (2012).
- Lau, M. C. C. *et al.* Pathogenesis of POLR1C-dependent type 3 Treacher Collins syndrome revealed by a zebrafish model. *Biochim. Biophys. Acta* **1862**, 1147–1158 (2016).
- Noack Watt, K. E., Achilleos, A., Neben, C. L., Merrill, A. E. & Trainor, P. A. The roles of RNA polymerase I and III subunits Polr1c and Polr1d in craniofacial development and in zebrafish models of Treacher Collins syndrome. *PLoS Genet.* **12**, e1006187 (2016).
- Valdez, B. C., Henning, D., Perumal, K. & Busch, H. RNA-unwinding and RNA-folding activities of RNA helicase II/Gu: two activities in separate domains of the same protein. *Eur. J. Biochem.* **250**, 800–807 (1997).
- Dixon, J., Brakebusch, C., Fässler, R. & Dixon, M. J. Increased levels of apoptosis in the prefrontal neural folds underlie the craniofacial disorder, Treacher Collins syndrome. *Hum. Mol. Genet.* **9**, 1473–1480 (2000).

13. Dixon, J. *et al.* Tcof1/Treacle is required for neural crest cell formation and proliferation deficiencies that cause craniofacial abnormalities. *Proc. Natl Acad. Sci. USA* **103**, 13403–13408 (2006).
14. Sloan, K. E. *et al.* The association of late-acting snoRNPs with human pre-ribosomal complexes requires the RNA helicase DDX21. *Nucleic Acids Res.* **43**, 553–564 (2015).
15. Gonzales, B. *et al.* The Treacher Collins syndrome (TCOF1) gene product is involved in pre-rRNA methylation. *Hum. Mol. Genet.* **14**, 2035–2043 (2005).
16. Jones, N. C. *et al.* Prevention of the neurocristopathy Treacher Collins syndrome through inhibition of p53 function. *Nat. Med.* **14**, 125–133 (2008).
17. Berkson, R. G. *et al.* Pilot screening programme for small molecule activators of p53. *Int. J. Cancer* **115**, 701–710 (2005).
18. Dolma, S., Lessnick, S. L., Hahn, W. C. & Stockwell, B. R. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. *Cancer Cell* **3**, 285–296 (2003).
19. Van Nostrand, J. L. *et al.* Inappropriate p53 activation during development induces features of CHARGE syndrome. *Nature* **514**, 228–232 (2014).
20. Rinon, A. *et al.* p53 coordinates cranial neural crest cell growth and epithelial-mesenchymal transition/delamination processes. *Development* **138**, 1827–1838 (2011).
21. Zhang, Z. *et al.* Assessment of hematopoietic failure due to Rpl11 deficiency in a zebrafish model of Diamond-Blackfan anemia by deep sequencing. *BMC Genomics* **14**, 896 (2013).
22. Sakai, D., Dixon, J., Achilleos, A., Dixon, M. & Trainor, P. A. Prevention of Treacher Collins syndrome craniofacial anomalies in mouse models via maternal antioxidant supplementation. *Nat. Commun.* **7**, 10328 (2016).
23. Muscarella, D. E., Ellison, E. L., Ruoff, B. M. & Vogt, V. M. Characterization of I-Ppo, an intron-encoded endonuclease that mediates homing of a group I intron in the ribosomal DNA of *Physarum polycephalum*. *Mol. Cell. Biol.* **10**, 3386–3396 (1990).
24. Flick, K. E., Jurica, M. S., Monnat, R. J. Jr & Stoddard, B. L. DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-Ppol. *Nature* **394**, 96–101 (1998).
25. Chailleux, C. *et al.* Quantifying DNA double-strand breaks induced by site-specific endonucleases in living cells by ligation-mediated purification. *Nat. Protocols* **9**, 517–528 (2014).
26. Song, C., Hotz-Wagenblatt, A., Voit, R. & Grummt, I. SIRT7 and the DEAD-box helicase DDX21 cooperate to resolve genomic R loops and safeguard genome stability. *Genes Dev.* <http://dx.doi.org/10.1101/gad.300624.117> (2017).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Stack and J. Wu for providing the endothelial and cardiomyocyte cells, J. Chen for *Xenopus* splicing morpholino validations, C. Santoriello and L. Zon for *ddx21* zebrafish morpholino, the Swanson Biotechnology Center at the Koch Institute for Integrative Cancer Research, especially E. Vasile for microscopy and A. Amsterdam for zebrafish work, and K. Cimplich and members of the Wysocka, Calo, and Chang laboratories for discussions. This work was supported by the Howard Hughes Medical Institute (J.W.), R01 GM112720 (J.W.), the March of Dimes Birth Defects Foundation (J.W.), March of Dimes Foundation grants 6-FY15-189 and RC35CA197591 (L.D.A.), Ludwig Foundation (J.W.), Stanford Medical Scientist Training Program and T32CA09302 (R.A.F.), the Helen Hay Whitney Foundation (E.C.), EMBO (ALTF 275-2015), the European Commission (LTFCONFUND2013, GA-2013-609409), and the Marie Curie Actions (A.Z.), Jane Coffin Childs Memorial Fund postdoctoral fellowship (M.E.B.), Stanford Graduate Student Fellowship (B.G.) and National Institutes of Health P50-HG007735 and R01-ES023168 (H.Y.C.).

Author Contributions J.W. supervised the project; E.C. conceived and designed the study; E.C. performed experiments with help from F.A. and J.L.; B.G. performed image analyses. E.C. and R.A.F. analysed ChIP-seq data; R.A.F. analysed the iCLIP data; F.A. and E.C. performed zebrafish experiments; M.E.B. performed mouse embryo dissections and immunostainings; A.Z. performed p53 *in situ* hybridization; J.L. and E.C. performed DNA damage experiments; T.S., L.D.A., and H.Y.C. provided advice on experimental designs, data analyses, and interpretation of the data; E.C. and J.W. wrote the manuscript with input from all co-authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.W. (wysocka@stanford.edu).

Reviewer Information *Nature* thanks D. Tollervey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Summary. HeLa cells were cultured in DMEM 10% FBS and maintained under standard tissue culture conditions unless otherwise specified. Mouse ES cells were cultured in 2i medium as previously described (see Methods). H9 human ES cells were maintained feeder-free and cultured in mTESR-1 (Stem Cell Technologies). U2OS cells were maintained in phenol-free DMEM supplemented with 10% Tet-free FBS. ChIP-seq and ChIP-qPCR analyses were conducted with commercially available antibodies described in the Methods sections. Nascent rRNA synthesis was performed using a Click-iT Nascent RNA Capture Kit (Thermo Fisher Scientific) following the manufacturer's instructions. For immunofluorescence studies, cells were fixed with 4% paraformaldehyde and/or methanol, unless otherwise specified, and stained with the indicated antibodies. Small-molecule treatments with DMSO, actinomycin D, or CX-5461 (Tocris) were done for 1 h, unless otherwise specified. Treatments with NSC146109 (Tocris), and DNAPK (Tocris, NU 7441) and ATM inhibitors (Tocris, KU 55933), were performed at the indicated time points in the figure legends. For annexin V staining, we used an Annexin V/Dead Cell Apoptosis Kit (Thermo Fisher Scientific) as recommended by the manufacturer. All animal work was done in accordance with the Stanford University Administrative Panel on Laboratory Animal Care. Zebrafish lines were housed in AAALAC-approved facilities and maintained according to protocols approved by the Massachusetts Institute of Technology Committee on Animal Care.

Cell lines. HeLa cells were obtained from American Type Culture Collection (ATCC) and grown under standard conditions in DMEM 10% FBS supplemented with antibiotics. U2OS cells were also obtained from ATCC and maintained in phenol-free DMEM supplemented with 10% Tet-free FBS. Human ES cells (H9 line) were obtained from WiCell, maintained and expanded feeder-free, and cultured in mTESR-1 (StemCell Technologies). Cells were passaged approximately 1:6 every 5–6 days with ReLeSR (StemCell Technologies) onto tissue culture dishes coated with growth-factor-reduced matrigel (BD Biosciences). Mouse ES cells were cultured in 2i medium as previously described²⁷. All cell lines used in this study were mycoplasma-free.

Vectors and plasmids. For generating GFP-TCOF1 and GFP-DDX21 expression vectors, gBlocks gene fragment (IDT) encoding the human TCOF1 and DDX21 coding region were synthesized and cloned into a GFP-expressing and tetracycline-inducible PiggyBac vector (SBI). The PiggyBac was stably integrated into pluripotent cells and several clones were expanded for further analyses. For generation of the cNCCs, we used our previously published protocol^{28,29} (described below). To express the transgenes, doxycycline was added to the medium at a concentration of $2.5 \mu\text{g ml}^{-1}$ for 24–48 h. For the generation of Tcof1 mutant ES cells and mouse ES cells, single guide RNAs (sgRNAs) were developed at <http://crispr.mit.edu/> and cloned into the PX458 vector (Addgene, 48138). Cells were transfected with Lipofectamine 2000 (Life Technologies), and sorted on the basis of GFP. The resulting clones were expanded and screened by PCR to search for Tcof1 loss-of-function alleles. I-PpoI construct (Addgene, 32565) was developed in M. Kastan's laboratory. AsiSI was synthesized as a gBlock gene fragment (IDT).

Differentiation of pluripotent cells into cNCCs and other lineages. Pluripotent lines were differentiated into cNCCs as previously described^{28,29}. In brief, human ES cells/mouse ES cells were incubated with 5 mg ml^{-1} collagenase (Life Technologies) for 1 h. Clusters of 100–200 cells were generated by manual pipetting and plated in a Petri dishes (BD Biosciences) containing cNCC differentiation medium (see methods in ref. 29 for details). The medium was changed every other day for 7 days or until the neuroepithelial spheres attached, before the induction and emigration of the cNCCs. To further purify the cNCCs, cells were dissociated with Accutase (Life Technologies) and passaged onto fibronectin-coated (Thermo Fisher Scientific) plates. cNCCs were then transitioned to maintenance media (see methods in ref. 29 for details). For propagating, cNCCs were passaged onto fibronectin-coated plates every three days. After two passages, they were transitioned to cNCC long-term maintenance medium (see methods in ref. 29 for details). For differentiation into smooth muscle cells, cNCCs were cultured in DMEM (Thermo Fisher Scientific) supplemented with 10% FBS (Omega). Differentiation of pluripotent cells into endothelial³⁰ cells and cardiomyocytes³¹ has been previously described by Wu's laboratory at Stanford University. Embryoid body differentiation was performed as previously described²⁷.

ChIP-qPCR and ChIP-seq. ChIP assays were performed as previously described⁵. In brief, cells were cross-linked with 1% formaldehyde for 10 min and quenched with glycine to a final concentration of 0.125 M for another 10 min. Chromatin was sonicated with a Bioruptor (Diagenode), cleared by centrifugation, and incubated overnight at 4°C with $7.5 \mu\text{g}$ of the desired antibodies: anti-DDX21 (Novus Biologicals BP100-1781), anti-TCOF1 (Novus Biologicals NBP1-86909). Immunocomplexes were immobilized with $100 \mu\text{l}$ of Protein G Dynal magnetic beads (Life Technologies) for 4 h at 4°C , followed by stringent washes and elution. Eluates were reverse cross-linked overnight at 65°C and de-proteinized with proteinase K at 56°C for 30 min. DNA was extracted with phenol–chloroform, followed by

ethanol precipitation. ChIP-seq libraries were prepared according to the NEBNext protocol and sequenced using Illumina HiSeq 2500. ChIP-qPCR analyses were performed in a Light Cycler 480II machine (Roche). ChIP-qPCR signals were calculated as the percentage of input. All primers used in qPCR analyses are shown in Supplementary Table 1.

ChIP-seq analyses. Sequences were mapped to the respective genomes using Bowtie2 and analysed by QuEST and MACS2. For QuEST, ChIP-seq peaks were determined using a kernel density estimate bandwidth of 30, a ChIP candidate threshold of $= 20$, a ChIP extension fold enrichment of 3, and a ChIP-to-background fold enrichment of 3. WIG files were generated with QuEST and used for visualization in the University of California, Santa Cruz (UCSC) Genome Browser. Average ChIP-seq signal profiles around the centre of DDX21 ChIP-seq peaks were generated with the Sitepro tool.

All genomic datasets have been deposited in the Gene Expression Omnibus under accession number GSE89420: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=abcjeymyvbkrap&acc=GSE89420>.

Mapping ChIP signal to rDNA locus. For mapping ChIP-seqs to the rDNA locus, we obtained the DNA consensus sequence of the 43 kb ribosomal locus NCBI (GeneBank accession number U13369.1). Using this unique 43 kb region, we used the Bowtie algorithm to map ChIP-seq reads with standard parameters used for mapping to the Hg19 human genome build (see methods in ref. 5 for details).

RNA extraction and RT-qPCR. RNA was isolated using TRIzol (Life Technologies) and an RNeasy mini kit (Qiagen) according to the manufacturer's protocol. All RNA samples were DNase treated in column (Qiagen). cDNA was generated using Sensifast (Bioline) according to the manufacturer's instructions. qPCR analyses were performed on a Light Cycler 480II (Roche). All primer sequences can be found in Supplementary Table 1.

Immunofluorescence. Cells were seeded into 24-well plates containing glass cover slips. For HeLa cells, mouse ES cells, and cNCCs, cover slips were coated with fibronectin and cultured for 16 h before treatment. After this, cells were treated for the indicated drugs (refer to the corresponding figure legends for drug concentration and timescale of the experiment). Cells were fixed in 4% PFA for 10 min at room temperature, washed three times for 5 min with PBS. For methanol fixation, after PFA fixation, cells were incubated with ice-cold methanol 2 min and washed twice for 5 min with PBS. Cells were permeabilized in PBS containing 0.3% (v/v) Triton X-100 for 5 min, and blocked overnight at 4°C in PBT buffer (PBS with 1% bovine serum albumin (BSA), 0.1% Triton X-100 (v/v), 0.05% sodium azide (w/v)). After blocking, coverslips were incubated in PBT with the corresponding antibody. For DDX21 (Novus Biologicals NBP1-83310), TCOF1 (Novus Biologicals NBP1-86909), $\gamma\text{H2A.X}$ (Abcam ab11174-50ug), and p53 (Vector Laboratories VP-P953), all antibodies were diluted 1:100 and incubated at room temperature for at least 2 h. Alexa Fluor 488, 568, and 647 secondary antibodies (1:1,000; Life Technologies) for 1 h. Cells were washed three times for 5 min with PBT, twice for 5 min with PBS, rinsed briefly with water, and mounted onto glass slides using VECTASHIELD mounting medium with DAPI. All images were taken and processed using a Zeiss LSM700 confocal microscope.

Western blots and co-immunoprecipitation. Nuclear extracts were prepared according to the Dignam and Roeder protocol³². For immunoprecipitations, extracts were incubated overnight with 3–5 μg of the desired antibody pre-bound to Protein G Sepharose (Pierce). In some cases, protein extracts were treated with RNaseA ($20 \mu\text{g ml}^{-1}$). Immunocomplexes were eluted in $2\times$ Laemmli buffer and resolved in a 4–20% pre-casted tris-glycine gel (Life Technologies). For western blots, the following antibodies were used according to the manufacturer's instructions: anti-GFP (Thermo Fisher Scientific A-6455), anti-TCOF1 (Novus Biologicals NBP1-86909 and Abnova H00006949-B01P), anti-Flag (Sigma-Aldrich), anti-DDX21 (Novus Biologicals NB100-1781), anti-p53 (Vector Laboratories VP-P953), anti-ACTIN (Abcam ab49900). All antibodies had been previously validated unless otherwise specified.

siRNA and antisense oligonucleotide knockdown. For knockdowns, HeLa cells were plated at a defined density (2.5×10^5) and transfected in DMEM supplemented with 5% FBS without antibiotics using RNAiMAX (Life Technologies) with 20 nM of the desired pools of siRNA. siRNA diced pools were generated in J.W.'s laboratory using recombinant *Giardia lamblia* Dicer.

Xenopus embryology. *X. laevis* embryos were staged in accordance with standard procedures³³. For knockdowns, morpholinos were injected into both blastomeres of two-cell-stage embryos at the indicated final concentrations. Embryo viability was scored throughout development to the tadpole stage. To make inferences about the observed phenotypes and determine the appropriate dosages, we conducted pilot experiments for each individual morpholino or *in vitro* transcribed RNA. We then determined the concentration at which the phenotypic variability was minimized within the sample and at the same time had minimal impact on the viability of the embryos. Once the dosages were determined, all embryos that survived were scored. No randomization of the sample or blinding were applied.

No statistical methods were used to predetermine sample size. The 5'-capped and 3'-tailed mRNAs were synthesized with a Message Machine Ultra kit (Ambion) and injected along with the indicated morpholino at the two-cell stage. For phenotypic analyses of craniofacial defects, stage 49 embryos were stained with alcian blue as previously described²⁸. The I-PpoI construct was deposited in Addgene (32565) by M. Kastan's laboratory. For injections, I-PpoI was amplified (see Supplementary Table 1 for oligonucleotides), followed by *in vitro* transcription using Message Machine Ultra kit (Ambion) and injected at the indicated concentrations.

Mouse husbandry. All mouse work was done in accordance with the Stanford University Administrative Panel on Laboratory Animal Care. Mice were maintained on a mixed 129/Sv-C57BL/6 background. The *Wnt1-cre* transgene and *Mdm2^{fllox}* alleles have been described previously^{34,35}. To obtain embryos, mice were mated overnight and the day a vaginal plug was observed was considered E0.5. Yolk sac DNA, extracted using the HotSHOT method³⁶, was used for genotyping. **Neural tube immunofluorescence.** Mouse embryos were fixed in 4% PFA overnight at 4°C, dehydrated in ethanol, paraffin embedded, and sectioned at 5 µm. For immunofluorescence, antigen retrieval was performed in citrate buffer (10 mM sodium citrate, pH 6.0, 0.05% Tween 20; 5 min at 'high pressure' in a pressure cooker) and slides were permeabilized with Tris-buffered saline containing 0.025% Triton X (TBS-TX). Blocking was performed for 1 h at room temperature with 10% goat serum and 1% BSA in TBS-TX. Primary antibody incubations were performed overnight at 4°C and secondary antibody incubations for 1 h at room temperature. The following antibodies were used and were diluted in 1% BSA in TBS-TX: mouse anti-p53 (1C12, Cell Signaling, 1:300), rabbit anti-DDX21 (NB100-1718, Novus, 1:100), fluorescein goat anti-rabbit (FI-1000, Vector Labs, 1:200), and Alexa Fluor 546 goat anti-mouse (A-11030, Thermo Fisher Scientific, 1:200).

Whole-mount *in situ* hybridization. In brief, embryos were collected in cold PBS and immediately fixed in 4% paraformaldehyde/PBS overnight at 4°C. Embryos were then dehydrated in methanol and bleached in a methanol/H₂O₂ (5:1) solution for 5 h at room temperature. Then embryos were rehydrated in PBS-T (0.1% Tween 20/PBS) before being incubated in a proteinase K solution (10 mg ml⁻¹ in PBS) for 15 min at room temperature and fixed in 0.2% glutaraldehyde/4% paraformaldehyde/PBS solution for 20 min at room temperature. Embryos were then incubated for 2 h at 65°C in the hybridization solution (composed of 2% Blocking Powder (Roche), 50% formamide (Sigma-Aldrich), 5× SSC (Sigma-Aldrich), 0.1% CHAPS (Sigma-Aldrich), 5 mM EDTA (Sigma-Aldrich), 0.1% Triton X-100 (Fisher) to which was added yeast tRNA (Sigma-Aldrich, 1 mg ml⁻¹) and Heparin (Sigma-Aldrich, 50 g ml⁻¹). After this step, the solution was replaced and the digoxigenin-labelled mRNA probe against p53 (Genebank, AB021961.1) added to the hybridization solution. Embryos were incubated in this solution overnight at 65°C. The next day, embryos were first washed in 50% formamide, 1× SSC, and 0.1% Tween 20, then they were washed several times in MABT (20 mM malic acid, 30 mM NaCl, 0.1% Tween 20) before being incubated for 2 h at room temperature in the blocking solution (2% Blocking Powder, 20% Inactivated Lamb Serum (Gibco) in MABT). Blocking solution was then diluted by half in MABT and the anti-digoxigenin antibody (Roche, 1/4,000) was added. Embryos were incubated in the antibody solution overnight at 4°C. Embryos were next washed for 2 days in MABT before proceeding to the revealing using BM Purple (Roche).

Image processing/analysis. All image analysis was performed using custom-written MATLAB scripts. For all images, the nuclear mask was generated by performing segmentation on DAPI images as follows. log-transformed images were convolved with a rotationally symmetric Laplacian of Gaussian filter, and objects were defined as contiguous pixels exceeding a threshold filter score calculated by Otsu's method. Specifically, for cNCC, a special segmentation method was used to de-clump the cells that clustered together and the poorly de-clumped cells were discarded for downstream analysis. In brief, a hierarchical watershed transform was performed where cells that had different levels of overlaps were separated by watershed with different parameters and the joint masks were defined to maximally separate the cells. Raw images were then subtracted with background signal defined by the average pixel intensity of the non-masked regions in the whole image, and all the following image analyses were done with background-subtracted images.

For nucleolus segmentation, a custom-written MATLAB function was implemented to reliably segment out the nucleoli regions from each individual cell by leveraging some or all main features of nucleoli in DAPI staining, including (1) low intensity of DAPI staining; (2) frequent appearance of DAPI high ridges surrounding nucleoli (separate nucleoli from Cajal bodies); (3) circularity (that nucleoli tend to be rounded). For each different cell type, four parameters were tuned to reach the maximum level of detection while limiting the false segmentation level.

For quantification of DDX21 shuttling, the ratio between the mean intensity of DDX21 signal within the defined nucleolar mask and the mean intensity of DDX21 signal within the nucleoplasmic mask (nuclear mask – nucleolar mask) was calculated for each individual cell.

For rRNA content quantification, the mean intensity of 5-ethynyl uridine staining within the defined nucleolar mask was calculated for each individual cell and a ratio-metric measurement performed to normalize the rRNA content to the mRNA content, where the nucleoli mean 5-ethynyl uridine intensity was divided by the nucleoplasmic 5-ethynyl uridine staining mean intensity for each individual cell.

For DNA damage analysis, a custom-written MATLAB function was implemented to segment out the γ H2A.X puncta. In brief, puncta were segmented by tophat filter and only those above the 75th percentile intensity of all puncta were called γ H2A.X foci. Each focus was then assigned with the nuclear mask identified with previous method and the number of foci/nuclei were then calculated. In particular, the perinucleolar γ H2A.X foci were defined by the foci identified above within a perinucleolar ring mask defined by a dilated nucleoli mask subtracting the nucleoli mask. See Supplementary Fig. 2 for details about the masks and segmentation methods.

Zebrafish embryology. Zebrafish *polr1c*^{+/-}, *polr1d*^{+/-}, and *rpl11*^{+/-} were obtained from the Zebrafish International Resource Centre. For crosses, we performed pilot experiments to determine that the published phenotypes were observed at the expected Mendelian ratios. Injection of *ddx21* morpholino or mRNA was performed on one-cell-stage zebrafish embryos at the indicated concentrations. Injected embryos were incubated at 28°C. To make inferences about the observed phenotypes and to determine the appropriate dosages, we conducted pilot experiments for *ddx21* morpholino or *in vitro* transcribed RNA. We then determined the concentration at which the phenotypic variability was minimized within the sample and at the same time had minimal impact on the viability of the embryos. Once the dosages were determined, all embryos that survived were scored. For cartilage staining with alcian blue, embryos were collected 5 days post-fertilization. Alcian blue staining was performed as described¹⁰. For haemoglobin staining with o-dianisidine and immunofluorescence, embryos were collected 24 h post-fertilization. o-Dianisidine staining was performed as described²¹. The sequence of *ddx21* morpholino was ATTCTGGGAGACTCTTTGACGGCAT. *DDX21* mRNA was transcribed from PCR products using a T7 mMessage Ultra Kit (Thermo Fisher Scientific) and purified using a MegaClear Kit (Thermo Fisher Scientific). For γ H2A.X immunofluorescence, fish of the indicated genotypes were harvested 24 h post-fertilization, fixed in 4% PFA diluted in PBS overnight, washed three times with 0.1% Tween-20 in PBS (PBS-T), and blocked with 1% BSA in PBS-T (PBS-TA) for 3 h. Embryos were stained in PBS-TA with zebrafish anti- γ H2A.X (GeneTex: GTX127340) overnight, washed three times with PBS-TA, and imaged using a Leica M205 stereomicroscope. No randomization of the sample or blinding were applied. No statistical methods were used to predetermine sample size.

iCLIP data analysis. FAST-iCLIP was performed³⁷ on HeLa cells by ultraviolet-crosslinking to a total of 0.35 J cm⁻². Whole-cell lysates were generated in iCLIP lysis buffer (50 mM HEPES, 200 mM NaCl, 1 mM EDTA, 10% glycerol, 0.1% NP-40, 0.2% Triton X-100, 0.5% N-lauroylsarcosine) and briefly sonicated using a probe-tip Branson sonicator to solubilize chromatin. Each iCLIP experiment was normalized for total protein amount, typically 1 mg, and partly digested with RNase I (Thermo Fisher Scientific, catalogue number AM2294) for 10 min at 37°C and quenched on ice. DDX21 was isolated with anti-DDX21 antibody (Novus Biologicals NB100-1718) for 8 h at 4°C on rotation. Samples were washed sequentially in 1 ml for 5 min each at 4°C: 2× high stringency buffer (15 mM Tris-HCl, pH 7.5, 5 mM EDTA, 2.5 mM EGTA, 1% Triton X-100, 1% sodium deoxycholate, 120 mM NaCl, 25 mM KCl), 1× high salt buffer (15 mM Tris-HCl pH 7.5, 5 mM EDTA, 2.5 mM EGTA, 1% Triton X-100, 1% sodium deoxycholate, 1 M NaCl), 1× NT2 buffer (50 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1 mM MgCl₂, 0.05% NP-40). After the NT2 wash, DDX21-bound RNA–protein complexes were dephosphorylated with T4 PNK (NEB, catalogue number M0210) for 30 min in an Eppendorf Thermomixer at 37°C, 15 s at 1,400 r.p.m., 90 s rest in a 30 µl reaction, pH 6.5, containing 10 units of T4 PNK, 0.1 µl SUPERase-IN, and 6 µl of PEG-400 (16.7% final). After 30 min, beads were rinsed once with NT2 buffer and 3' end ligated with T4 RNA Ligase 1 (NEB, catalogue number M0204) overnight in an Eppendorf Thermomixer at 16°C, 15 s at 1,400 r.p.m., 90 s rest in a 30 µl reaction containing 10 units of T4 RNA Ligase, 1 pmole pre-adenylated-DNA-adaptor, 0.1 µl SUPERase-IN, and 6 µl of PEG400 (16.7% final). The following day, samples were again rinsed with NT2 buffer and 5' radiolabelled by adding 1 µl of T4 PNK, 0.5 µl g32-ATP (Perkin Elmer), 2 µl 10× T4 PNK buffer, 0.5 µl SUPERase-IN, and 16 µl of water for 15 min at 37°C. To this reaction, 1 µl of 100 mM DTT and 6 µl of 4× LDS Buffer (Thermo Fisher Scientific) were added, samples headed to 75°C for 10 min, and released RNA–protein complexes were separated on SDS–PAGE using NuPAGE 4–12% Bis-Tris Gels (1.0 mm × 12 wells) at 180 V for 45 min. Resolved ribonucleoprotein complexes were wet-transferred to nitrocellulose at 400 mA for 60 min at 4°C.

RNA was recovered and processed for library preparation as in the irCLIP protocol³⁸. Membranes were cut into approximately 0.5 × 1 mm narrow strips

that easily came to rest in the bottom of a siliconized 1.5 ml Eppendorf tube. To each tube, we added 0.2 ml of proteinase K reaction buffer (100 mM Tris, pH 7.5, 50 mM NaCl, 1 mM EDTA, 0.2% SDS) and 10 µl of proteinase K (Thermo Fisher Scientific, catalogue number AM2546). The reaction was incubated for 60 min at 50 °C in an Eppendorf Thermomixer. Next, 200 µl of saturated phenol–chloroform, pH 6.7, was added to each tube and incubated for 10 min at 37 °C in an Eppendorf Thermomixer, 1,400 r.p.m. Tubes were briefly centrifuged and the entire contents transferred to a 2 ml Heavy Phase Lock Gel (5Prime, catalogue number 2302830). After 2 min centrifugation at more than 13,000 r.p.m., the aqueous layer was re-extracted with 1 ml of chloroform (invert tube ten times to mix; do not vortex, pipet or shake) in the same 2-ml Phase Lock Gel tube and centrifuged for 2 min at more than 13,000 r.p.m. The aqueous layer was then transferred to a new 2-ml Heavy Phase Lock Gel tube and extracted again with an additional 1 ml of chloroform. After 2-min centrifugation at more than 13,000 r.p.m., the aqueous layer was transferred to a siliconized 1.5-ml Eppendorf tube and precipitated overnight at –20 °C by addition of 10 µl 5 M NaCl, 3 µl Linear Polyacrylamide (Thermo Fisher Scientific, catalogue number AM9520), and 0.8 ml ethanol.

cDNA synthesis primers were purchased from IDT: cDNA-barcode1 (6 bp TruSeq barcode in bold type): /5phos/WWNNNNXXXXNNNNNTACCCTT CGCTTCACACACAAG/iSp18/GGATCC/iSp18/TACTGAACCGC. P3short (cDNA elution oligonucleotide): CTGAACCGCTCTTCCGATCT. PCR1 primers, P3tall: GCATTCCTGCTGAACCGCTCTTCCGATCT; P6tall: TTTCCC CTGTGTGTGAAGCGAAGGGTA. PCR2 primers (PAGE purified), P3solexa: CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCCTGCTGAACCG CTCTTCCGATCT; P6solexa: AATGATACGGCGACACCGAGATCTACACT CTTTCCCTTGTGTGTGAAGCGAAGGGTA. P6 sequencing primer (for Illumina sequencing): CACTCTTCCCTTGTGTGTGAAGCGAAGGGTA.

RNA fragments were pelleted at more than 13,000 r.p.m. for 45 min at 4 °C, washed once with 1 ml of ice-cold 75% ethanol and air dried. Pellets were resuspended in 12 µl water. RNA (12 µl) was mixed with 1 µl of 1 µM cDNA and 1 µl of 10 mM dNTPs and heated to 70 °C for 10 min, then rapidly cooled to 4 °C. Six microlitres of cDNA Master Mix (4 µl 5× SSIV buffer, 1 µl 100 mM DTT, 1 µl SSIV) was added to the annealed RNA and incubated for 30 min at 55 °C. cDNA:RNA hybrids were captured by addition of 5 µl of MyOne Streptavidin C1 Dynabeads (Thermo Fisher Scientific, catalogue number 65001) that had been rinsed and suspended in 30 µl of biotin-IP buffer (100 mM Tris, pH 7.5, 1 M NaCl, 1 mM EDTA, 0.1% Tween), and end-over-end rotation for 30 min at room temperature. Beads were placed on a 96-well magnet and washed sequentially with 0.1 ml of Biotin-IP buffer and PBS. Beads were resuspended in 10 µl of cDNA elution/ RNA degradation buffer (8.25 µl water, 1 µl of 1 µM P3short oligo, and 0.75 µl of 50 mM MnCl₂) and placed in a thermocycler with the following program: 5 min at 95 °C, 1 min at 75 °C, ramp 0.1 °C s^{–1} to 60 °C forever. After 15 min, tubes were removed and mixed with 5 µl of Circligase-II reaction buffer (3.3 µl water, 1.5 µl 10× Circligase-II buffer, and 0.2 µl of Circligase-II, Epicentre, catalogue number CL9021K). cDNA was circularized in a thermocycler for 1.5 h at 60 °C. cDNA was captured by addition of 30 µl of Ampure XP beads (Beckman Coulter, catalogue number A63880), 75 µl of isopropanol, and incubation for 15 min (the solution was remixed after 7.5 min). Beads were washed once with 80% ethanol, dried for 5 min, and resuspended in 14 µl of water. For maximal elution, tubes were placed in a 95 °C thermocycler for 2 min and immediately transferred to a 96-well magnet. The 14 µl eluate was transferred to a new 0.2 ml PCR tube containing 15 µl of 2× Phusion HF-PCR Master Mix (NEB, catalogue number M0531), 0.5 µl of 30 µM P3/P6 PCR1 oligo mix, and 0.5 µl of 15× SYBR Green I (Thermo Fisher Scientific, catalogue number S7563). The tubes were then placed in a Stratagene MX3000P qPCR machine with the following program: 98 °C for 2 min, 15 cycles of 98 °C for 15 s, 65 °C for 30 s, 72 °C for 30 s, with data acquisition set to the 72 °C extension. PCR1 reactions were then subjected to one round of magnetic bead size selection by addition of 4.5 µl of isopropanol, 54 µl of Ampure XP beads, and incubation for

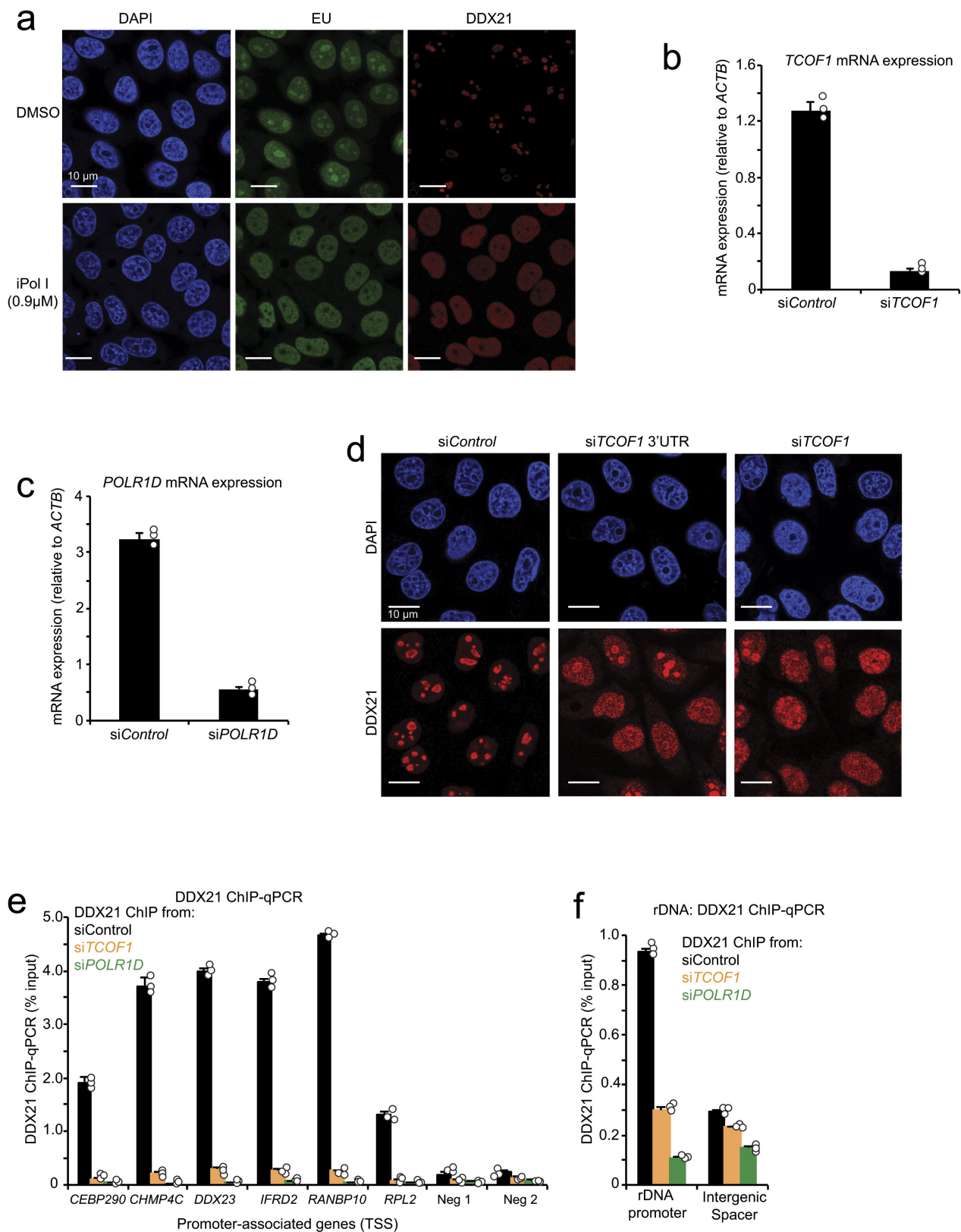
10 min. Beads were washed once with 80% ethanol, dried for 5 min, and eluted in 10 µl of water. PCR1 products were subjected to a second round of size selection by addition of 1.5 µl of isopropanol, 18 µl of Ampure XP beads, and incubation for 10 min. Beads were washed once with 80% ethanol, dried for 5 min, and eluted in 10 µl 500 nM P3solexa/P6solexa oligonucleotide mix. Ten microlitres of 2× Phusion HF-PCR Master was added to each tube and placed in a thermocycler with the following program: 98 °C for 2 min, three cycles of 98 °C for 15 s, 65 °C for 30 s, 72 °C for 30 s. Final libraries were purified by addition of 36 µl of Ampure XP beads and incubation for 5 min. Beads were washed twice with 70% ethanol, dried for 5 min, and eluted in 20 µl of water. One microlitre of each library was quantitated by HS-DNA Bioanalyzer.

Samples were sent for deep sequencing on an Illumina NextSeq machine for single-end 75-bp cycle runs. FAST-iCLIP data were processed using the FAST-iCLIP analysis pipeline (<https://github.com/ChangLab/FAST-iCLIP>). PCR duplicates were removed using unique molecular identifiers in the RT primer region. Adaptor and barcode sequences were trimmed, and reads were mapped stepwise to repetitive and non-repetitive genomes. Specific parameters used were as follows: –f 18 (trims 17 nt from the 5′ end of the read), –l 15 (includes all reads longer than 15 nt), –bm 25 (minimum MAPQ score from Bowtie2 of 25 is required for repeat element mapping), –sr 0.08 (STAR mismatch-per-base ratio; 0.08 corresponds to 2 mismatches per 25 bases), and –tr 2,3 (repetitive genome) and –tn 2,3 (non-repetitive genome) RT stop intersection (n,m ; where n = replicate number and m = number of unique RT stops required per n replicates). Using the –tr/tn 2,3 parameters, a minimum of six RT stops were required to support any single nucleotide identified as crosslinking site.

Code availability. All custom codes are available upon request from the corresponding author.

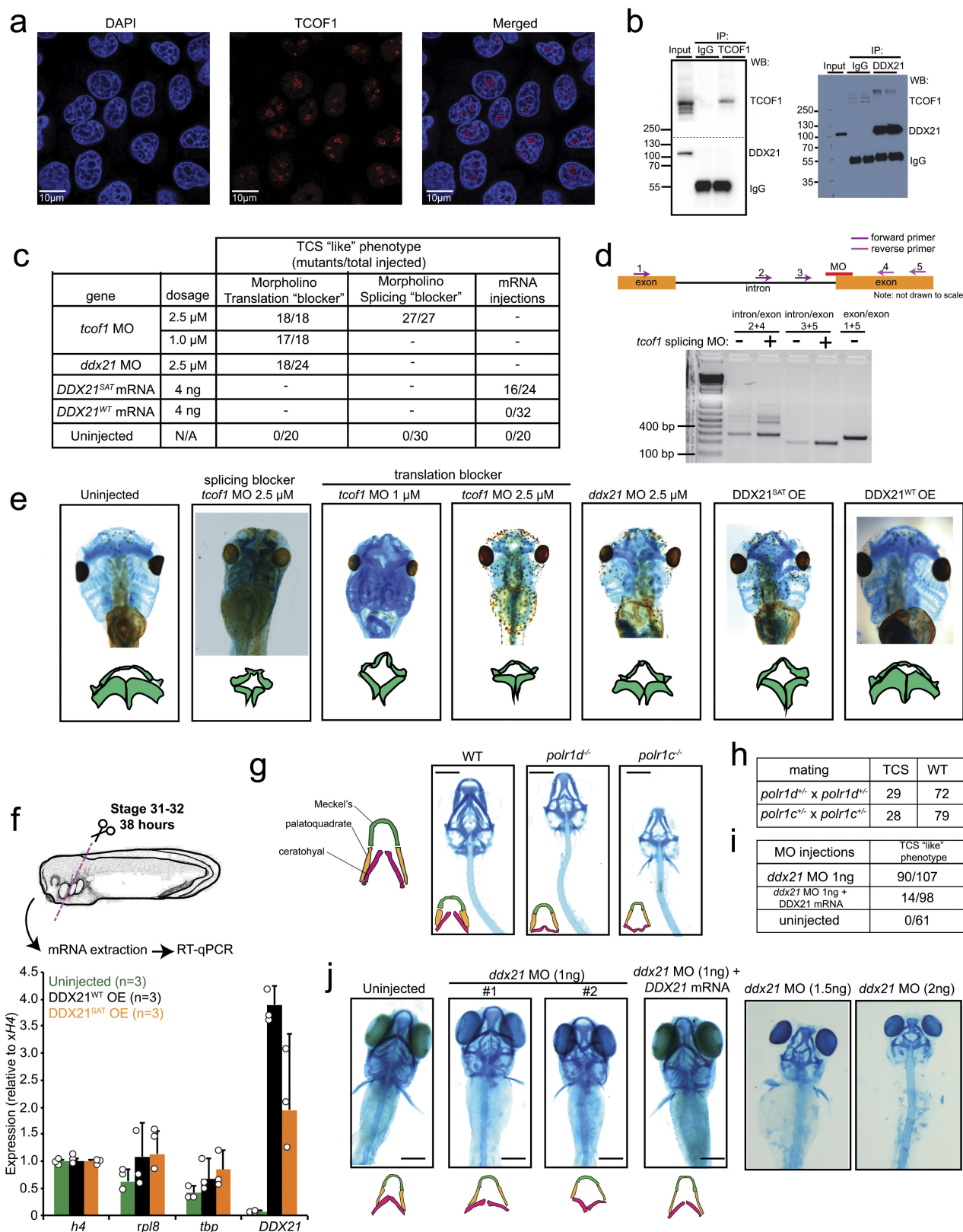
Data availability. Data that support the findings of this study have been deposited in the Gene Expression Omnibus under accession number GSE89420. Source Data for Figs 1–4 and Extended Data Figs 8 and 9 have been provided. Source Data for western blots shown in Fig. 4 and Extended Data Figs 2–5, 7 and 9 have been provided in Supplementary Fig. 1. All other relevant data that support the findings of this study are available from the corresponding author upon reasonable request.

27. Buecker, C. *et al.* Reorganization of enhancer patterns in transition from naive to primed pluripotency. *Cell Stem Cell* **14**, 838–853 (2014).
28. Bajpai, R. *et al.* CHD7 cooperates with PBAF to control multipotent neural crest formation. *Nature* **463**, 958–962 (2010).
29. Prescott, S. L. *et al.* Enhancer divergence and cis-regulatory evolution in the human and chimp neural crest. *Cell* **163**, 68–83 (2015).
30. Hu, S. *et al.* Effects of cellular origin on differentiation of human induced pluripotent stem cell-derived endothelial cells. *JCI Insight* **1**, e85558 (2016).
31. Burrage, P. W. *et al.* Chemically defined generation of human cardiomyocytes. *Nat. Methods* **11**, 855–860 (2014).
32. Carey, M. F., Peterson, C. L. & Smale, S. T. Dignam and Roeder nuclear extract preparation. *Cold Spring Harb. Protoc.* 2009, <http://dx.doi.org/10.1101/pdb.prot5330> (2009).
33. Nieuwkoop, P. D. & Faber, J. (eds) *Normal Table of Xenopus laevis (Daudin): A Systematical and Chronological Survey of the Development from the Fertilized Egg Till the End of Metamorphosis* (Garland, 1994).
34. Grier, J. D., Yan, W. & Lozano, G. Conditional allele of mdm2 which encodes a p53 inhibitor. *Genesis* **32**, 145–147 (2002).
35. Danielian, P. S., Muccino, D., Rowitch, D. H., Michael, S. K. & McMahon, A. P. Modification of gene activity in mouse embryos in utero by a tamoxifen-inducible form of Cre recombinase. *Curr. Biol.* **8**, 1323–1326 (1998).
36. Truett, G. E. *et al.* Preparation of PCR-quality mouse genomic DNA with hot sodium hydroxide and tris (HotSHOT). *Biotechniques* **29**, 52–54 (2000).
37. Flynn, R. A. *et al.* Dissecting noncoding and pathogen RNA–protein interactomes. *RNA* **21**, 135–143 (2015).
38. Zarnegar, B. J. *et al.* irCLIP platform for efficient characterization of protein–RNA interactions. *Nat. Methods* **13**, 489–492 (2016).



Extended Data Figure 1 | DDX21 subnuclear localization is sensitive to perturbations in the rRNA synthesis. **a**, Representative immunofluorescence depicting DDX21 localization and 5-ethynyl uridine (EU) incorporation in HeLa cells treated with DMSO or iPol I from $n = 3$ biologically independent experiments. **b**, **c**, siRNA pools were developed against *TCOF1* or *POLR1D* and transfected into HeLa cells. qPCR was used to determine knockdown efficiency. **d**, An additional pool of siRNAs targeting *TCOF1* 3' UTR was generated and transfected into HeLa cells,

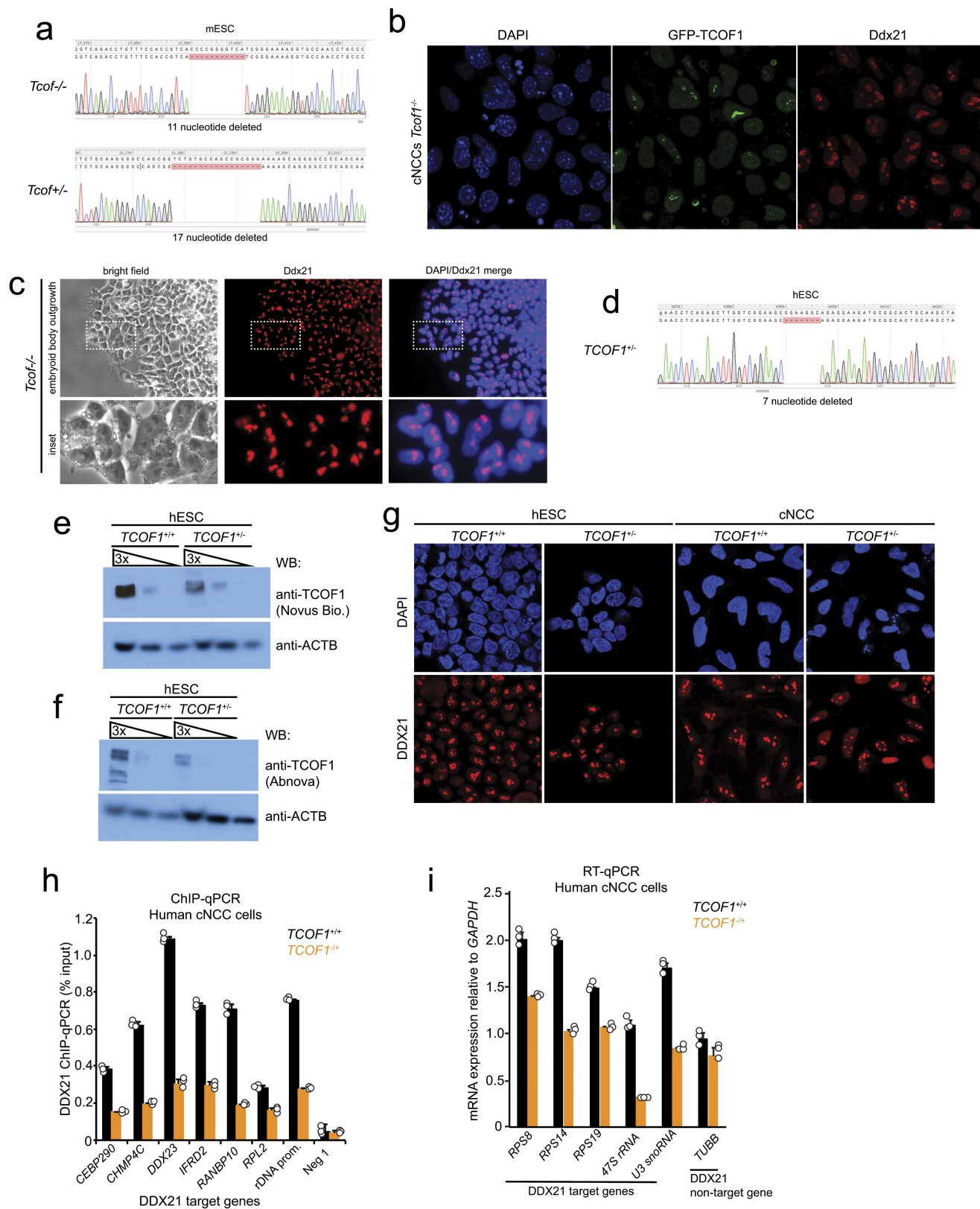
followed by immunofluorescence to determine DDX21 localization upon *TCOF1* knockdown. Shown are representative images from $n = 3$ biologically independent experiments. **e**, **f**, ChIP-qPCR of DDX21 binding at target gene promoters (**e**) and the rDNA locus (**f**) upon knockdown of either *TCOF1* (siTCOF1) or *POLR1D* (siPOLR1D). For **b**, **c**, **e** and **f**, bars represent the average $n = 3$ biologically independent experiments; error bars, s.e.m.



Extended Data Figure 2 | See next page for caption.

Extended Data Figure 2 | DDX21 knockdown phenocopies TCS-associated perturbations in *X. laevis* and zebrafish. **a**, Representative immunofluorescence images showing strong nucleolar localization signal for TCOF1 in HeLa cells from $n = 3$ biologically independent experiments. **b**, Immunoprecipitation of either GFP-tagged TCOF1 (GFP-TCOF1) or DDX21, followed by western blotting with the indicated antibodies. $n = 2$ biologically independent experiments. **c**, Table showing the quantification of injected *Xenopus* embryos with the indicated morpholinos (MO) or *in vitro* transcribed mRNAs. **d**, Efficiency of *tcof1* splicing morpholino was determined by PCR. $n = 10$ injected embryos. **e**, Representative images of stage 49 *Xenopus* embryo cartilage stainings with alcian blue. Traces of the mandibular and hyoid streams are shown for clarity. Embryos were collected from $n = 3$ biologically independent experiments. **f**, Stage 2 embryos were injected with *in vitro* transcribed mRNAs encoding wild-type or catalytically defective DDX21. Total mRNA was extracted

at stage 31, followed by qPCR to determine the expression levels of injected mRNAs in the anterior part of the embryo (see schematics for details). Bars represent the average of $n = 3$ independent experiments; error bars, s.e.m. **g**, Representative images of 5-day-old wild-type (WT), *polr1d*^{-/-}, and *polr1c*^{-/-} zebrafish embryo cartilage stained with alcian blue from $n = 3$ independent matings. **h**, Table showing the quantification of *polr1d*^{-/-} and *polr1c*^{-/-} crosses. **i**, Table showing quantification of zebrafish embryos injected with the indicated morpholino or combination of morpholino and mRNA. **j**, Representative images of 5-day-old zebrafish embryo cartilage stained with alcian blue after injection of *ddx21* morpholino at the indicated dosages or *ddx21* morpholino and *in vitro* transcribed human DDX21 mRNA. Traces of the ceratohyal, platoquadrate, and Meckel's cartilage are shown for clarity. $n = 3$ biologically independent sets of injections.

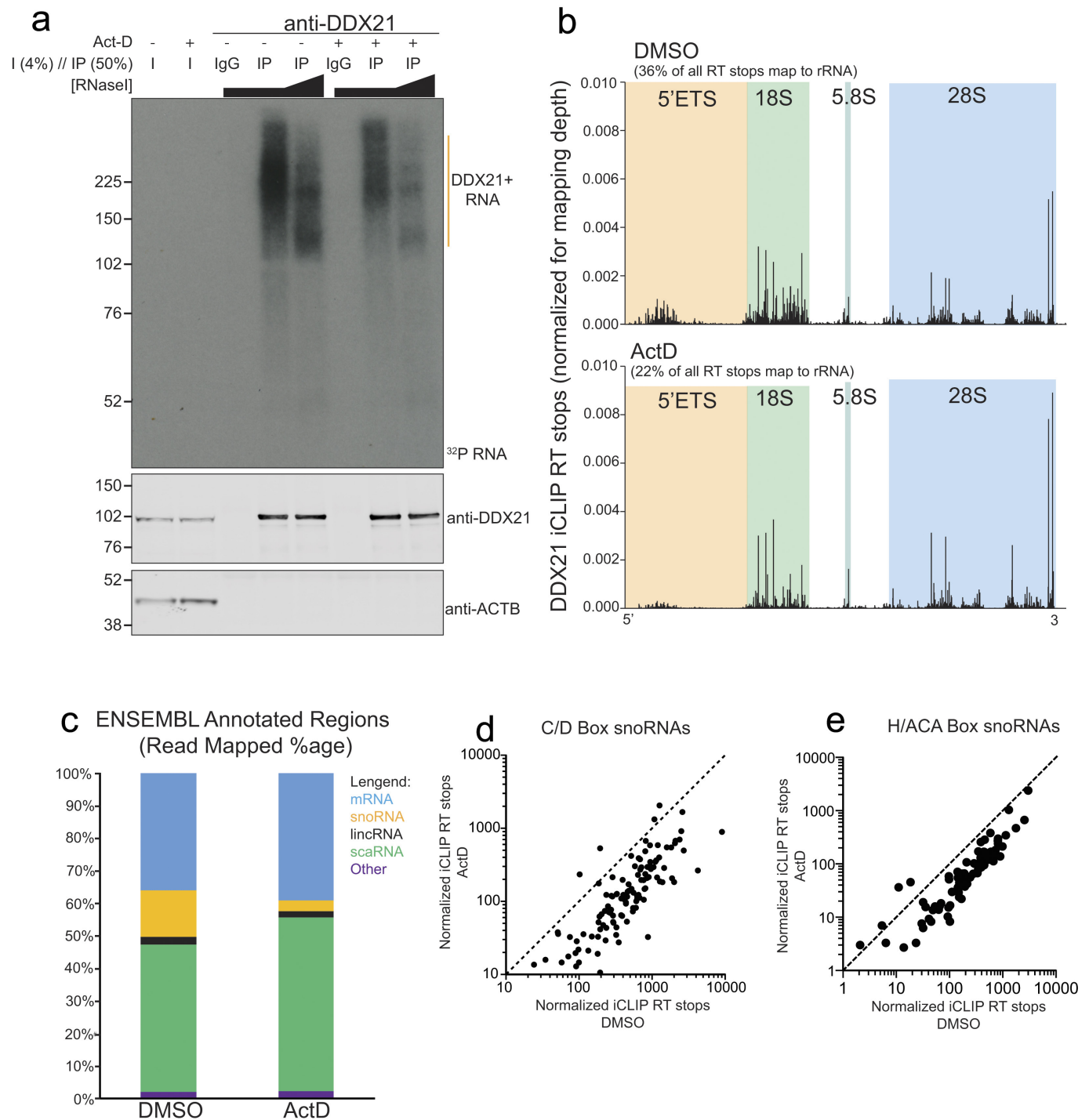


Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Generation of an *in vitro* model of TCS.

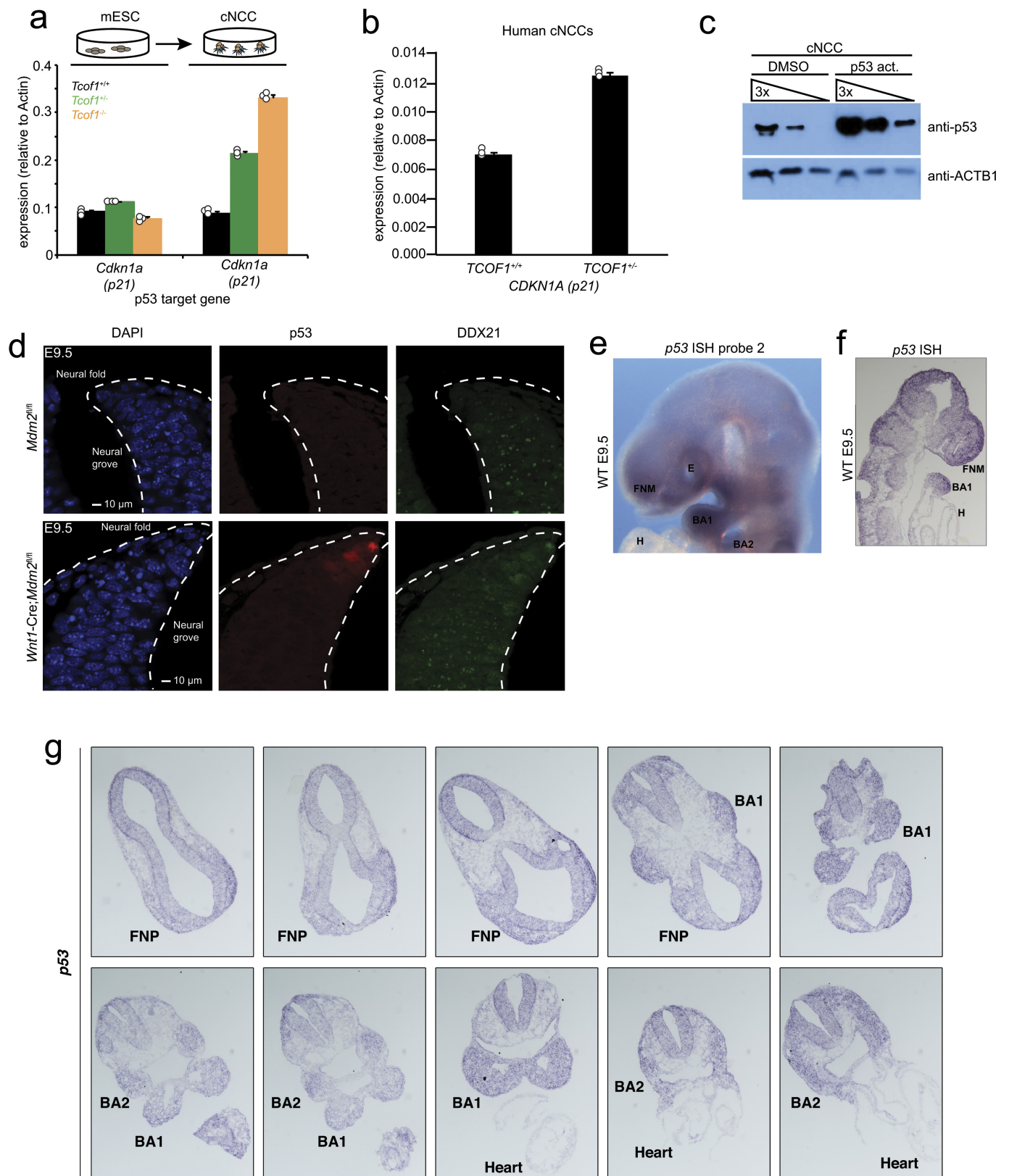
a, Mouse ES cells were co-transfected with CRISPR–Cas9 and sgRNAs targeting the *Tcof1* locus. Targeted mouse ES cells were single-cell cloned and screened for loss-of-function mutations in *Tcof1*. Clones of the indicated genotypes were selected for this study. **b**, Overexpression of an exogenous, but stably integrated human GFP-tagged TCOF1 construct in mouse cNCCs rescued DDX21 localization defects as determined by immunofluorescence (quantifications are shown in Fig. 2c). Shown are representative images from $n = 3$ biologically independent experiments. **c**, Mouse ES cells were differentiated into embryoid bodies. Embryoid body outgrowth explants were further grown in culture and stained with antibodies for Ddx21. Shown are representative images from $n = 4$ biologically independent experiments. **d**, Human H9 ES cells were co-transfected with CRISPR–Cas9 and sgRNAs targeting the *TCOF1* locus.

Targeted ES cells were cloned and screened for loss-of-function mutations in *TCOF1*; unlike mouse ES cells, we did not recover homozygous mutant alleles for *TCOF1* in human cells. The indicated genotype was selected for this study. **e**, **f**, Two different commercially available antibodies raised against TCOF1 were used to confirm the heterozygosity of *TCOF1*^{+/-} ES cells. Shown are representative western blots from $n = 2$ biologically independent experiments. **g**, Representative immunofluorescence images showing DDX21 localization in both wild-type and *TCOF1*^{+/-} human ES cells and cNCCs from $n = 3$ biologically independent experiments. **h**, ChIP–qPCR analysis in human cNCCs sampling DDX21 genomic occupancy at a representative panel of DDX21 target promoters and at the rDNA promoter. **i**, qPCR analyses of DDX21-regulated Pol I and Pol II transcribed ribosomal genes. For **h**, **i**, bars represent the average of $n = 3$ biologically independent experiments; error bars, s.e.m.



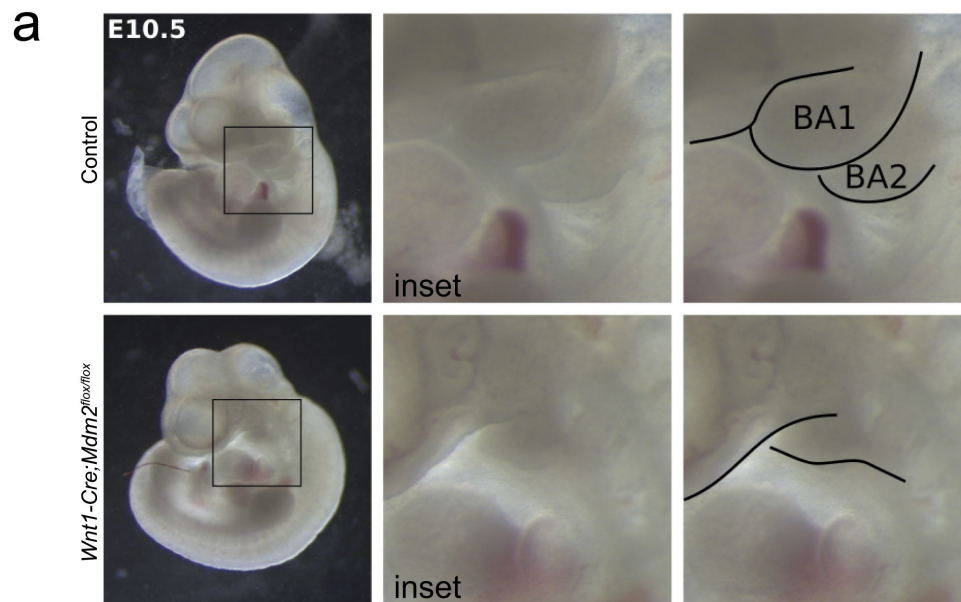
Extended Data Figure 4 | Pol I inhibition impairs the ability of DDX21 to associate with the 5' external transcribed spacer (ETS) and the snoRNAs. **a**, DDX21 iCLIP ^{32}P -autoradiogram and western blots from control (DMSO) and Pol I inhibited cells. For Pol I inhibition, we used low levels of actinomycin D (ActD; 50 ng ml^{-1}). Samples were loaded with constant input lysate amounts. **b**, DDX21 iCLIP reads mapped to the

transcribed region of the rDNA. The 5' external transcribed spacer and the mature portions of the 18S, 5.8S, and 28S rRNAs are highlighted. **c**, Distribution of ENSEMBL annotated regions for DDX21-bound RNAs in both DMSO and actinomycin D conditions. **d**, **e**, Scatter plot analysis of normalized iCLIP reverse transcription stops on individual C/D or H/ACA snoRNAs. iCLIP results are from $n = 2$ biological replicates.



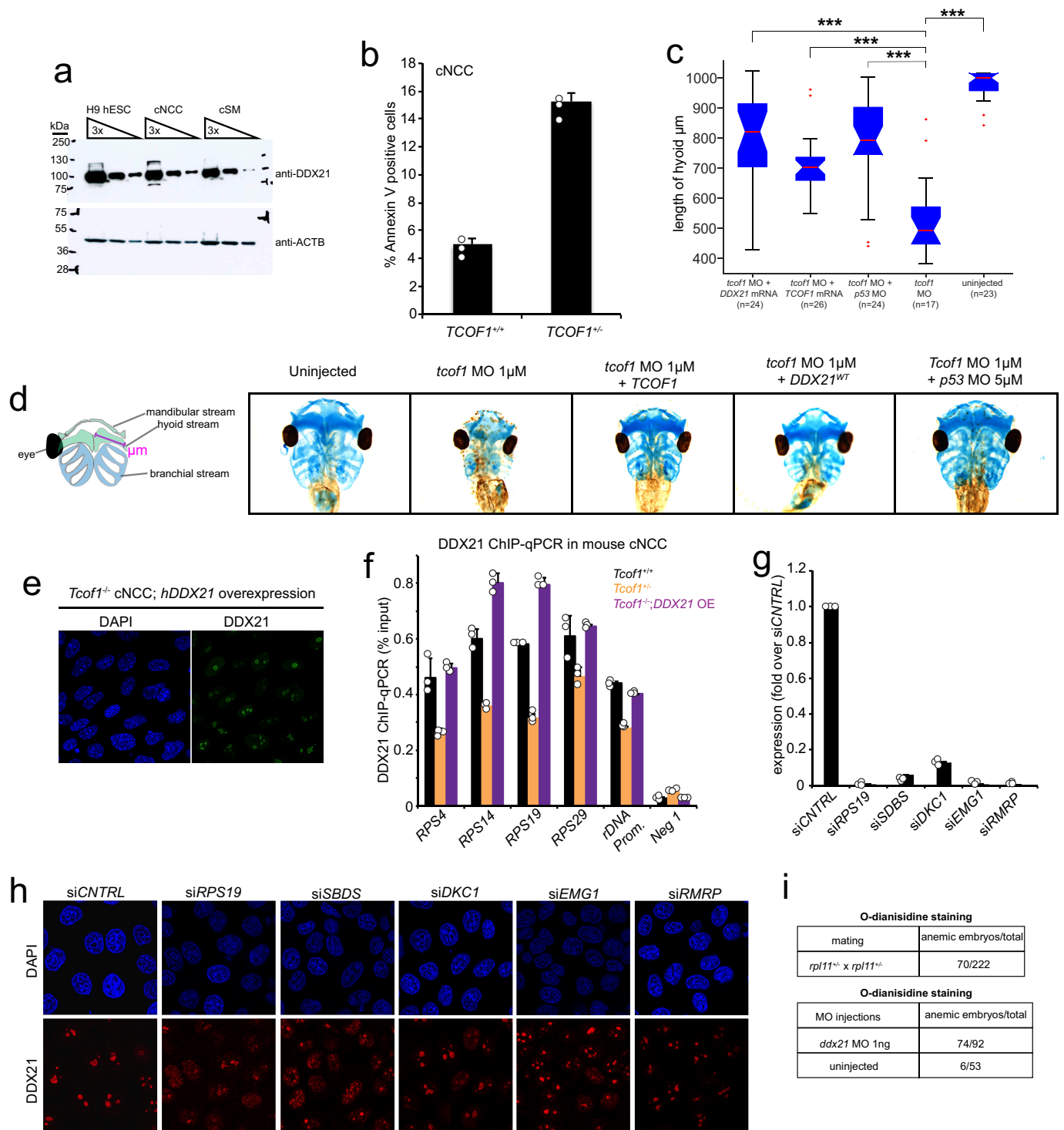
Extended Data Figure 5 | Protein p53 is activated in TCS cNCCs and its mRNA is highly expressed in cNCCs *in vivo*. **a, b**, qPCR analyses of the p53-target gene *Cdkn1a* in mouse ES cells and cNCCs (**a**) and human cNCCs (**b**) of indicated genotypes. Bars represent the average of $n = 3$ biologically independent experiments; error bars, s.e.m. **c**, Human cNCCs were treated with NSC146109 for 12 h, followed by western blotting with antibodies raised against p53. Shown is a representative western blot from $n = 3$ biologically independent experiments. **d**, Immunofluorescence

staining of p53 and DDX21 in sections from the dorsal neural tube of *Mdm2*^{fl/fl} (control; top) and *Wnt1-cre;Mdm2*^{fl/fl} (bottom) E9.5 mouse embryos. Dotted lines outline the neural fold. $n = 5$ animals per genotype. **e**, Representative picture of whole-mount *in situ* hybridization of E9.5 embryos with a probe recognizing endogenous *p53* mRNA. $n = 4$ animals. **f, g**, Representative images of *p53* *in situ* hybridization on tissue sections of the frontonasal prominence and first and second pharyngeal arches of E9.5 mouse embryos. $n = 2$ independent animals.



Extended Data Figure 6 | Hyper-activation of p53 in cNCCs renders pharyngeal arches hypoplastic. Representative images of wild-type and *Wnt1-cre;Mdm2^{fl/fl}* E10.5 embryos. Whole-embryo pictures (left) and

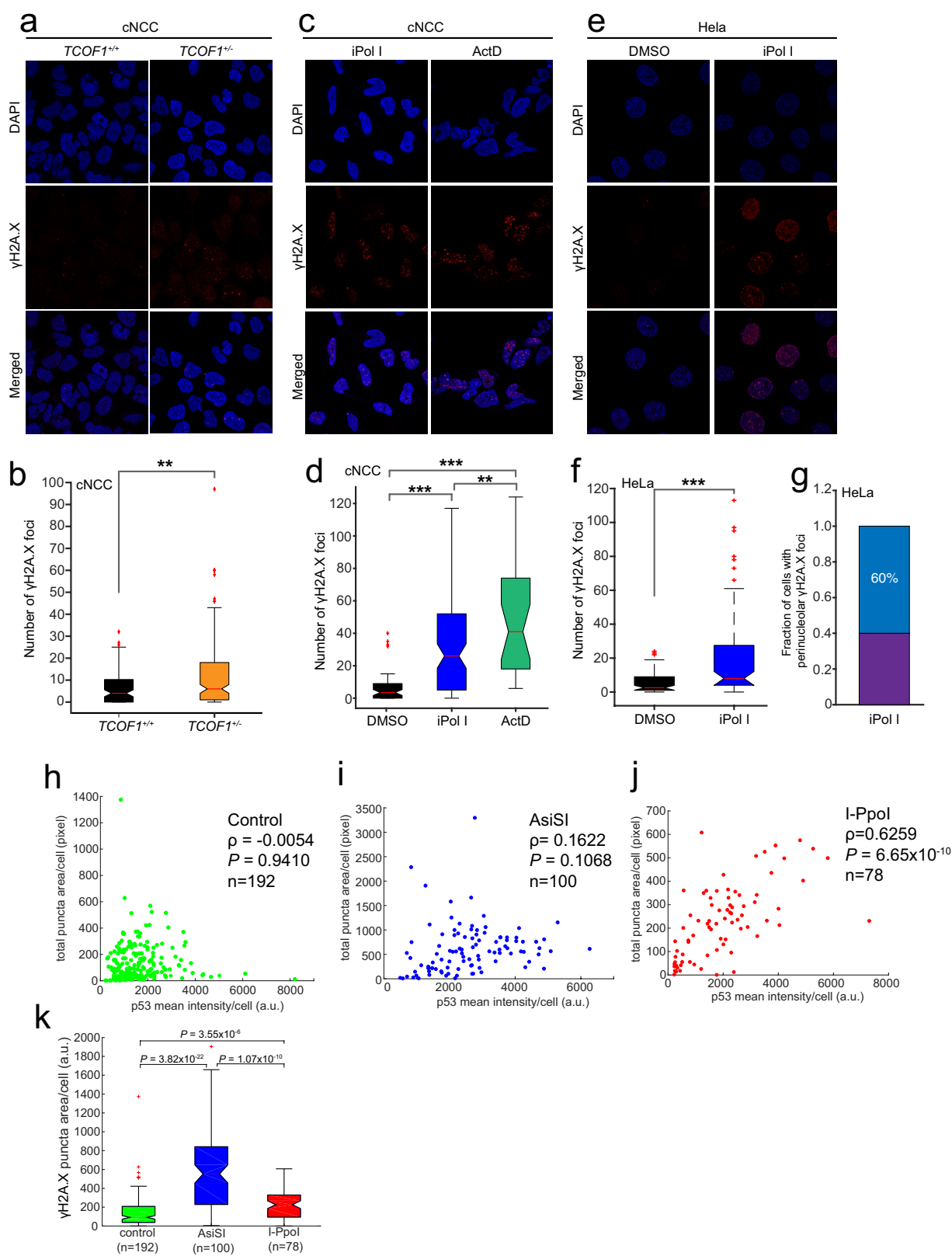
insets (middle) depicting the location of the first and second pharyngeal arches. Right panel shows traces of first and second pharyngeal arches for clarity. $n = 8$ animals per genotype.



Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | DDX21 overexpression rescues TCS and its function is deregulated by knockdown of other ribosomopathy-associated genes. **a**, Representative western blot for DDX21 in different cell types from $n = 3$ biologically independent experiments. **b**, FACS analyses to determine the sensitivity of $TCOF1^{+/-}$ cNCC to p53-mediated apoptosis. cNCCs were treated with NSC146109 for 4–6 h (note that this time point is significantly shorter than the one used in Fig. 3f and g). Apoptosis was quantified by FACS of annexin V staining. Bars represent the average of $n = 3$ independent experiments; error bars, s.e.m. **c**, Quantification of *Xenopus* craniofacial development rescue experiments by measuring the length of the hyoid stream upon overexpression of *TCOF1*, *DDX21*, or *p53* knockdown. Embryos were collected from $n = 3$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. *** $P < 0.001$, two-sided Wilcoxon–Mann–Whitney test. **d**, Rescue of TCS-associated craniofacial malformations in *Xenopus* by injection of the embryos with the indicated *in vitro* transcribed mRNAs and/or morpholinos (quantification is shown in c).

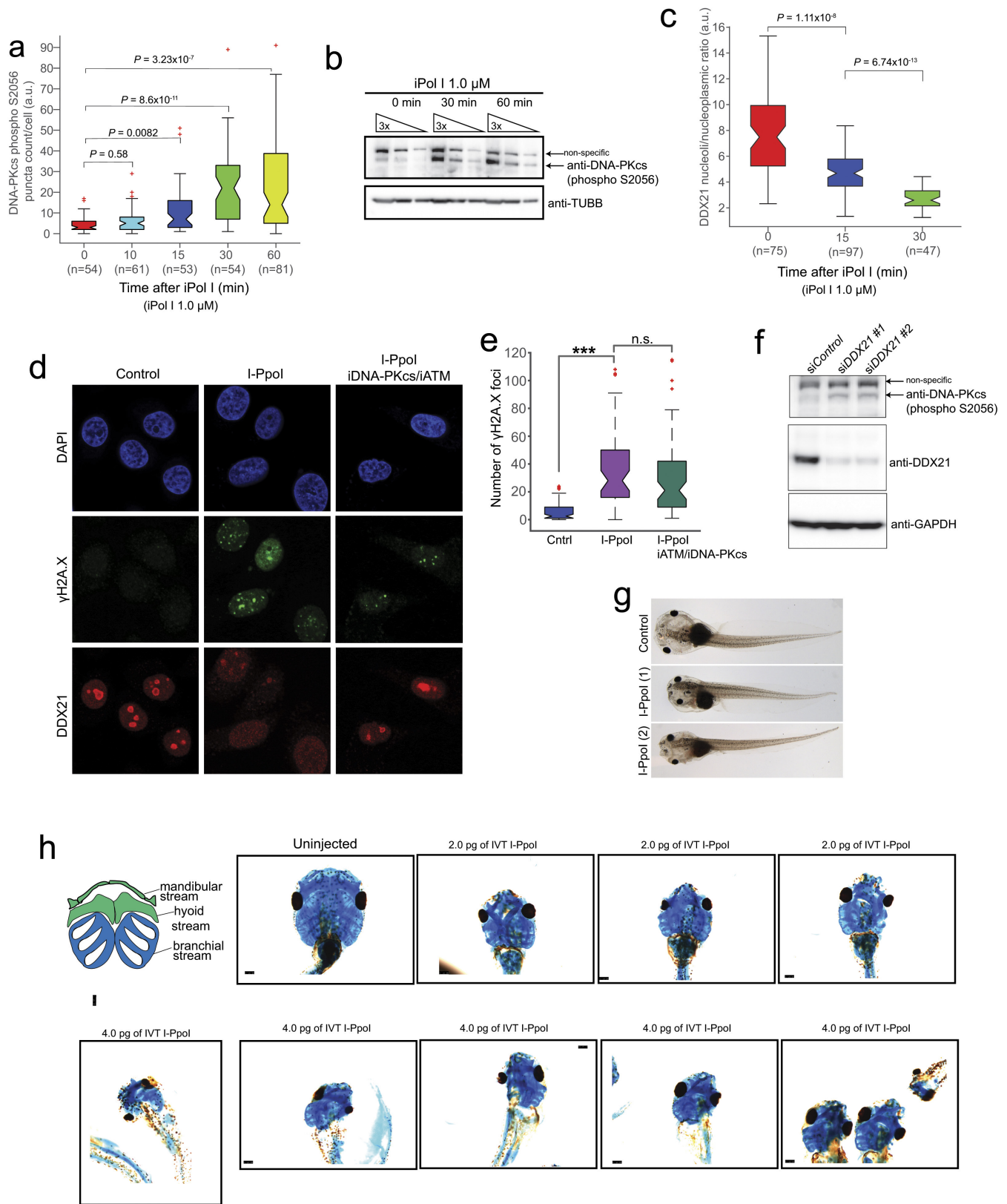
e, Representative immunofluorescence images of mouse *Tcof1*^{-/-} cNCCs upon overexpression of human GFP-tagged DDX21. $n = 3$ biologically independent experiments. **f**, ChIP–qPCR analysis, in mouse cNCCs sampling Ddx21 genomic occupancy, at a representative panel of Ddx21 target promoters and at the rDNA locus. $n = 3$ biologically independent experiments. Bars represent the average of $n = 3$ independent experiments; error bars, s.e.m. **g**, siRNA pools against a subset of ribosomopathy-associated genes were transected into HeLa cells. qPCR was used to determine the efficiency of the knockdowns. Bars represent the average of $n = 3$ independent experiments; error bars, s.e.m. **h**, Representative immunofluorescence images showing DDX21 localization changes in HeLa cells transfected with the indicated pools of siRNAs (quantification is on Fig. 3i). $n = 3$ biologically independent experiments. **i**, Tables quantifying the number of embryos stained for haemoglobin with o-dianisidine for the indicated genotypes. In the case of *rpl1* zebrafish, embryos were collected from three independent matings. For DDX21, three independent batches of embryos were injected and stained.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Inhibition of Pol I results in DNA damage in a subset of cells. **a**, Representative immunofluorescence images of wild-type and *TCOF1*^{+/-} cNCCs stained with an antibody against γ H2A.X; quantification is shown in **b**. **c**, Representative immunofluorescence images of DNA-damaged wild-type cNCCs stained with an antibody against γ H2A.X after 1 h treatment with iPol I or actinomycin D (ActD); quantification is shown in **d**. **e**, Representative immunofluorescence images of DNA-damaged HeLa cells stained with an antibody against γ H2A.X after 1 h treatment with iPol I; quantification is shown in **f**. For **a–f**, cells were collected from $n = 3$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. *** $P < 0.001$, two-sided

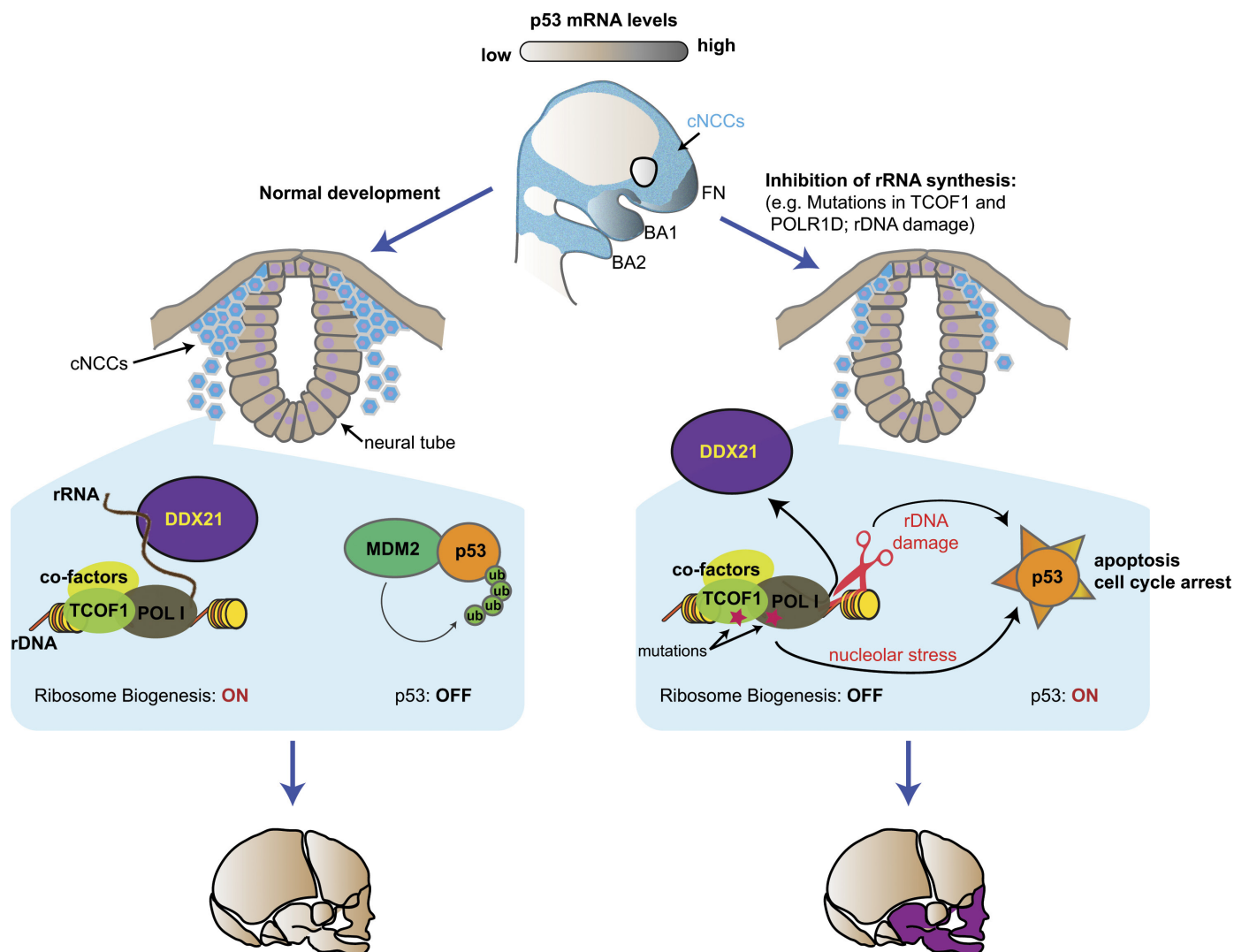
Wilcoxon–Mann–Whitney test. **g**, Fraction of DNA-damaged HeLa cells displaying perinucleolar γ H2A.X signal after 1 h incubation with iPol I. Cells were collected from $n = 3$ biologically independent experiments. **h–j**, Single-cell correlation plots of p53 activation and γ H2A.X signal in control and cells expressing either AsiSI or I-PpoI. Cells were collected from $n = 4$ biologically independent experiments. ρ , Pearson correlation coefficient. P , two-sided Wilcoxon–Mann–Whitney test. **k**, Single-cell quantification of γ H2A.X signal in control and cells expressing either AsiSI or I-PpoI. Cells were collected from $n = 4$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. P , two-sided Wilcoxon–Mann–Whitney test.



Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | rDNA damage impairs DDX21 functions and causes craniofacial deformities. **a, b**, Time course of phosphorylated DNA-PKcs as measured by auto-phosphorylation of the kinase in Ser2056 (S2056) upon iPol I treatment. Cells were collected from $n = 3$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. P , two-sided Wilcoxon–Mann–Whitney test. **c**, Time course of DDX21 exclusion from the nucleolus to the nucleoplasm upon iPol I treatment. Cells were collected from $n = 3$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. P , two-sided Wilcoxon–Mann–Whitney test. **d**, Representative immunofluorescence images of HeLa cells transfected with *in vitro* transcribed *I-PpoI* and treated or not with inhibitors for ATM (iATM) and DNAPK (iDPK). After treatment, cells were stained with antibodies against DDX21 and γ H2A.X.

e, Box plot quantifying the number of γ H2A.X foci of HeLa cells transfected with *I-PpoI* and treated or not with iATM and iDPK. Cells were collected from $n = 3$ biologically independent experiments. Boxes represent median value and 25th and 75th percentiles, whiskers are minimum to maximum, crosses are outliers. $***P < 0.001$, two-sided Wilcoxon–Mann–Whitney test, not significant (NS). **f**, Western blot showing DNA-PKcs by auto-phosphorylation of S2056 upon knockdown of DDX21. Two different siRNAs were used in this experiment (see Methods for details). $n = 3$ biologically independent experiments. **g**, Representative bright-field images of stage 49 *Xenopus* embryos either uninjected or injected with *in vitro* transcribed *I-PpoI* and **(h)** alcian blue stainings of *Xenopus* cranial cartilage from embryos injected with the indicated doses of *in vitro* transcribed *I-PpoI*. Embryos were collected from $n = 4$ biologically independent injections.



Extended Data Figure 10 | Model explaining cNCC-type selective effects of nucleolar dysfunction and rDNA damage in TCS. cNCCs express high levels of *p53* mRNA, but during normal development *p53* is under post-transcriptional control of its E3 ligase, Mdm2. Upon nucleolar stress and/or rDNA damage, activation of *p53* and loss of DDX21 from

chromatin result in apoptosis of a subset of cNCCs. This diminishes the population of cNCCs that can be allocated to the lower face, leading to malformations of the developing craniofacial structures. Thus, factors whose perturbations may ultimately induce defects in rRNA synthesis and rDNA damage are likely to be associated with craniofacial malformations.

A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria

Kathryn M. Kauffman¹, Fatima A. Hussain¹, Joy Yang¹, Philip Arevalo¹, Julia M. Brown^{2†}, William K. Chang², David VanInsberghe¹, Joseph Elsherbini¹, Radhey S. Sharma^{1‡}, Michael B. Cutler¹, Libusha Kelly² & Martin F. Polz¹

The most abundant viruses on Earth are thought to be double-stranded DNA (dsDNA) viruses that infect bacteria¹. However, tailed bacterial dsDNA viruses (*Caudovirales*), which dominate sequence and culture collections, are not representative of the environmental diversity of viruses^{2,3}. In fact, non-tailed viruses often dominate ocean samples numerically⁴, raising the fundamental question of the nature of these viruses. Here we characterize a group of marine dsDNA non-tailed viruses with short 10-kb genomes isolated during a study that quantified the diversity of viruses infecting *Vibrionaceae* bacteria. These viruses, which we propose to name the *Autolykiviridae*, represent a novel family within the ancient lineage of double jelly roll (DJR) capsid viruses. Ecologically, members of the *Autolykiviridae* have a broad host range, killing on average 34 hosts in four *Vibrio* species, in contrast to tailed viruses which kill on average only two hosts in one species. Biochemical and physical characterization of autolykiviruses reveals multiple virion features that cause systematic loss of DJR viruses in sequencing and culture-based studies, and we describe simple procedural adjustments to recover them. We identify DJR viruses in the genomes of diverse major bacterial and archaeal phyla, and in marine water column and sediment metagenomes, and find that their diversity greatly exceeds the diversity that is currently captured by the three recognized families of such viruses. Overall, these data suggest that viruses of the non-tailed dsDNA DJR lineage are important but often overlooked predators of bacteria and archaea that impose fundamentally different predation and gene transfer regimes on microbial systems than on tailed viruses, which form the basis of all environmental models of bacteria–virus interactions.

The dsDNA viruses consist of two ancient major lineages, both of which are proposed to have evolved from viruses that infect bacteria, and both include members that infect all three domains of life^{5–8}. These two lineages emerged from ancestors with distinct folds in their major capsid proteins, the HK97 fold⁹ and the DJR fold¹⁰, and among the dsDNA bacterial viruses, these two groups are recognizable as ‘tailed’ and ‘non-tailed’ viruses, respectively. However, despite the DJR being the second most common capsid fold among all viral taxa¹¹, with the single jelly roll fold being the most common, bacterial DJR viruses are essentially missing from culture and sequence collections, which are instead dominated by HK97-lineage tailed viruses¹². Whereas there are 215 described genera of tailed viruses¹³, with 1,993 *Caudovirales* genomes in the NCBI RefSeq database (as of 3 October 2017)¹⁴, there are only three described genera of non-tailed DJR bacterial and archaeal viruses, and 8 NCBI RefSeq genomes. Notably, only one of these sequenced DJR non-tailed viruses, the corticovirus PM2, which was isolated 50 years ago, is of marine origin¹⁵. This is particularly puzzling, given that electron microscopy-based surveys have revealed that non-tailed viruses comprise 51–92% of viruses observed in global surface oceans^{4,16} and dsDNA viruses are thought to represent the majority of

marine viruses¹⁷, suggesting that non-tailed dsDNA viruses should be abundant. Directed efforts have led to the discovery that non-tailed RNA viruses that infect eukaryotes can be abundant¹⁸, and that non-tailed single-stranded DNA (ssDNA) viruses that infect bacteria¹⁹ are also diverse, although with a low abundance²⁰, in marine systems. However, it remains unresolved whether non-tailed dsDNA viruses, such as those in the ancient and diverse DJR capsid lineage, are contributors to the enigmatic non-tailed majority of viruses that dominate the global ocean.

In a large survey of viruses that infect the ubiquitous marine bacterial family *Vibrionaceae*, we recovered a diverse collection of non-tailed viruses from a quantitative assay that exposed 1,334 *Vibrionaceae* isolates to concentrates of co-occurring viruses (Methods). We used a quantitative isolation approach that enabled the capture of all viruses

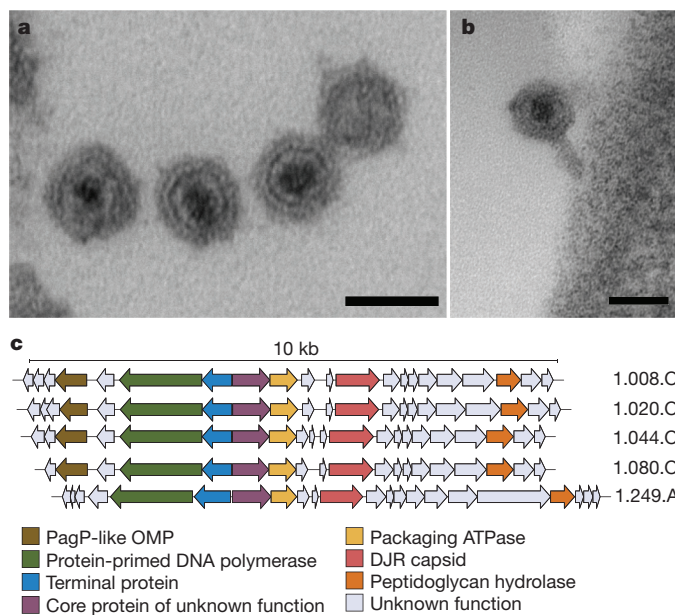


Figure 1 | *Autolykiviridae* is a new family of non-tailed dsDNA viruses in the DJR capsid lineage. **a**, Thin-section electron microscopy of autolykivirus plaques shows non-tailed virions with inner cores similar to those of the lipid-bilayer-containing non-tailed corticovirus PM2 (see Methods for experimental details and references). **b**, Rare virions show a tectivirus-like tail-tube-like structure when adjacent to cell membrane. Scale bars, 50 nm. **c**, Alignment of five genomes representing autolykivirus diversity, open reading frames are represented by block arrows. The linear 10-kb autolykivirus genomes have inverted terminal repeats and are shorter than those of the tailed viruses described here, which range from 21.7–348.9 kb (median = 47 kb) and encompass the range of tailed *Vibrio* virus genomes in GenBank.

¹Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Department of Systems and Computational Biology, Albert Einstein College of Medicine, Bronx, New York 10461, USA. [†]Present addresses: Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544, USA (J.M.B.); Department of Environmental Studies, Bioresources & Environmental Biotechnology Laboratory, University of Delhi, Delhi 110007, India (R.S.S.).

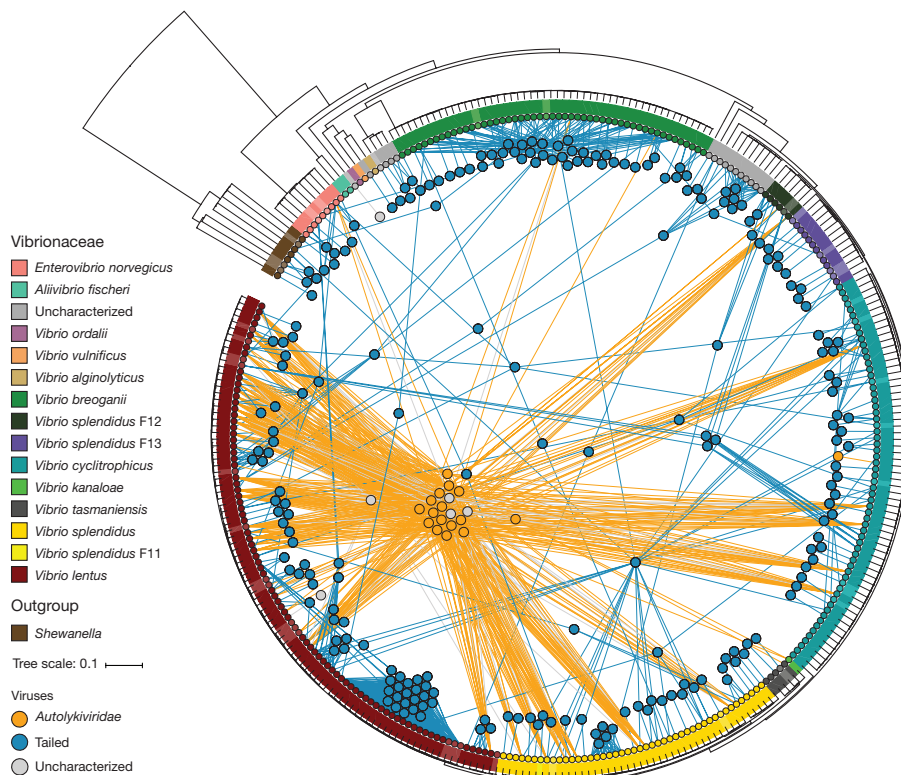


Figure 2 | Autolykiviruses dominate the lytic viral infection network of marine *Vibrio*. Inverted phylogenetic tree showing relationships among all 318 bacterial strains assayed based on concatenated alignments of *hsp60* and ribosomal protein genes, and using a partitioned model in RAxML³¹ to allow placement of 40 strains for which only the *hsp60* gene sequence

was available. Isolates are predominantly non-clonal. Leaves represent bacterial isolates coloured by species. Nodes represent 247 viruses described as *Autolykiviridae* ($n = 17$), tailed ($n = 224$) or uncharacterized ($n = 6$; no genome sequence). The edges represent infections coloured by viral type.

capable of lytic growth and colony (plaque) formation, mixing viral concentrates from co-occurring water samples and incubating them in solid-phase agar overlay for two weeks. Sequencing of 241 viruses that were randomly selected from each of 239 different plaque-positive hosts indicated that 18 of these viruses were a novel type that had small genomes (approximately 10 kb). Electron microscopy revealed that these viruses were non-tailed (Fig. 1a), although we also observed rare virions that showed tail-tube-like structures when in contact with cell membranes (Fig. 1b and Extended Data Fig. 1), consistent with known formation of such tubes during infection by other non-tailed viruses, including the dsDNA *Tectiviridae* (PRD1)²¹ and the ssDNA *Microviridae* (PhiX174)²². Notably, the capsid size of a representative member (mean \pm s.d. diameter, 49 ± 2 nm; Methods) of these novel viruses was closely aligned to the most abundant viral capsid size (mean \pm s.d. 54 ± 12 nm) observed in the surface ocean by electron microscopy⁴. This size is similar to the size of the only described non-tailed marine dsDNA and RNA isolates of bacterial viruses, PM2 and 06 N-58P, respectively, which both have 60-nm diameter capsids⁴, but is different from the size of the six described non-tailed ssDNA isolates of bacterial viruses, which have bimodal capsid diameter distributions centred around 31 nm and 72 nm (ref. 4). These observations suggest that these new viral isolates are representatives of the non-tailed viral majority.

Genome sequences and phylogenetic analyses of the non-tailed dsDNA *Vibrio* viruses show that they represent a new family of bacterial viruses, which we propose to name *Autolykiviridae*, in reference to Autolykos, a character in Greek mythology notable for being difficult to catch. Genome alignments of autolykivirus isolates reveal that they are diverse at the nucleotide level (Extended Data Fig. 2a), with whole-genome nucleotide identity as low as 31% (Extended Data Fig. 2b), yet display high synteny overall—sharing a core of six of their approximately 20 proteins, with additional proteins shared among subsets of

the isolates (Fig. 1c, Extended Data Fig. 3 and Supplementary Data 1). Phylogenetic analyses reveal that members of the *Autolykiviridae* are most closely related to the corticovirus PM2 in their major capsid protein (21–25% amino acid identity, Extended Data Fig. 4a, b), are poorly resolved in their packaging ATPase (12–16% amino acid identity to *Corticoviridae* and *Turriviridae* viruses, Extended Data Fig. 4c, d) and are most closely related to members of the *Tectiviridae* in their protein-primed DNA polymerase (36–37% amino acid identity, Extended Data Fig. 4e, f). The high sequence divergence of autolykiviruses in these core genes, in addition to their divergent phylogenetic association with previously described virus families, supports their identification as a family-level lineage.

To characterize the potential ecological impact of autolykiviruses, we conducted a large-scale host-range assay, and found that they commonly killed hosts in multiple species, whereas the majority of tailed viruses killed only few and closely related hosts. We used a collection of *Vibrionaceae* viruses that were isolated by quantitative direct plating and tested infectivity of 241 viruses on 318 bacterial isolates. We found that the autolykiviruses were disproportionate contributors to lysis, responsible for 38% of killings although representing only 7% of all tested viruses (Fig. 2). Notably, despite the high genomic diversity of the autolykiviruses and of the hosts they infect, these viruses share extensively overlapping host-range profiles (Methods and Extended Data Fig. 5). This pattern is similar to that observed for members of the *Tectiviridae*, which infect hosts in multiple Gram-negative genera in a plasmid-dependent manner²³. The finding that the autolykiviruses more commonly infect diverse species within a genus than tailed viruses suggests that these two groups may have fundamentally different impacts on microbial community ecology and evolution.

Biochemical and phenotypic characterization of members of the *Autolykiviridae* revealed several properties that make them subject to systematic loss in studies of viral diversity. Firstly, we found that

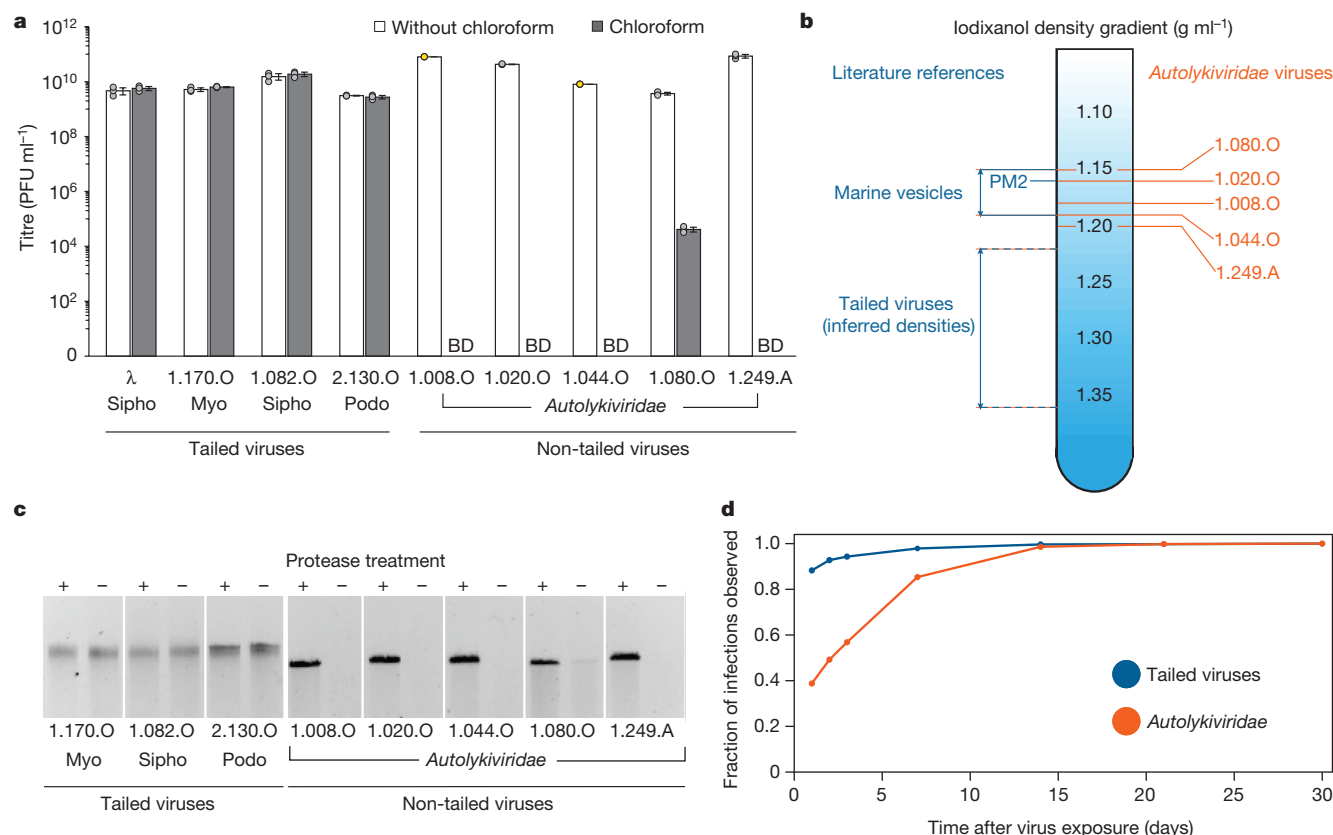


Figure 3 | Recovery of autolykiviruses is subject to multiple methodological biases. **a**, Comparison of chloroform sensitivity of tailed viruses and representative autolykiviruses (Extended Data Fig. 2), measured by plaque-forming assay after chloroform exposure. Data are mean \pm s.d. of three independent replicates, data points in yellow represent lower-bound values. BD, below the detection limit of 199 plaque forming units (PFU) per ml. **b**, Buoyant density of autolykiviruses in iodixanol, in relation to previously reported densities for the lipid-containing non-tailed corticovirus PM2 and marine (outer membrane) vesicles (solid lines) in iodixanol, and inferred range of caesium-chloride-targeted tailed

viruses (dashed lines) on the basis of linear extrapolation from PM2 (see Methods). **c**, Comparison of tailed virus and autolykivirus genome recovery with and without protease treatment. Protease-treated sample loading volumes normalized to 50 ng, equal volumes of untreated partner samples in adjacent lanes. The cropped gel image is representative of three independent experiments (gel source data are shown in Supplementary Fig. 1). **d**, Comparison of the cumulative proportion of observed tailed virus and autolykivirus killing over time. Infections ($n = 498$ and 844 , autolykiviruses and tailed viruses, respectively) assayed as drop-spot clearings in large-scale host-range assay.

chloroform (Fig. 3a), a reagent that is commonly added to viral preparations to kill contaminating bacterial cells²⁴, reduced infectivity of autolykiviruses to below the level of detection. Secondly, we observed lower buoyant densities for autolykiviruses than those inferred for tailed viruses, probably owing to the presence of a lipid bilayer within their capsid, placing them outside the range that is commonly targeted in density gradient-based preparations of bacterial viruses from environmental samples^{24,25} (Fig. 3b and Methods). Thirdly, owing to the presence of covalently bound proteins that alter DNA partitioning, the genomes of autolykiviruses require treatment with protease to enable efficient DNA extraction; this is not a standard component of extraction protocols targeting tailed viruses (Fig. 3c). Additional features, such as time to detection and decay rate, may also contribute to recovery bias (Methods, Fig. 3d and Extended Data Fig. 6). That *Autolykiviridae*-like viruses have not been definitively described for *Vibrio*, which have served as a major model of host–virus interactions and have been used to isolate viruses for nearly 100 years²⁶, suggests that the impact of these biases is severe and that related viruses that infect a diverse range of other bacteria are also likely to be systematically lost as a result of the same biases. We therefore suggest that, except for studies that specifically target subsets of viruses, viral concentrates for isolation and metagenomics are prepared: (1) without chloroform, (2) without density gradients and (3) with protease treatment during extraction.

Using combined cultivation and bioinformatic approaches, we show that DJR elements also exist as actively mobilizing prophages and episomes in *Vibrio*. Genome-integrated elements that have previously

been identified as widespread putative corticovirus-like prophages²⁷ are active and naturally excise to produce nuclease-protected extracellular particles (Extended Data Fig. 7a, b). Furthermore, a set of broad host-range plasmids that have previously been identified as non-transmissible²⁸ encode DJR capsid proteins and associated packaging ATPases and are thus also DJR elements (Extended Data Fig. 7c, d). These findings suggest that DJR-encoding mobile elements that have been identified in cellular sequence databases, either as plasmids or integrated prophages, contribute to the pool of environmental non-tailed viruses.

We next investigated how diverse DJR elements are among bacteria and archaea, as well as in the marine environment. Considering the paucity and high divergence of reference sequences, we used a two-phase iterative hidden Markov model-based search approach to first generate a broad panel of DJR capsid sequences associated with putative prophages of bacterial and archaeal genomes, and to then search nine cellular and viral metagenomes that represent marine sediment- and water-column-derived samples (Methods and Extended Data Table 1), as well as NCBI environmental metagenomes.

Our searches reveal that the diversity and host associations of DJR viruses far exceeds the level that is currently recognized, as putative DJR prophage capsids were identified in 13 bacterial and archaeal phyla and metagenomic sequences, suggesting the existence of at least 13 additional novel lineages with unknown hosts. Whereas DJR viruses and prophages had previously been shown to infect two archaeal⁶ and two bacterial phyla^{23,27}, the first phase of our search revealed the presence

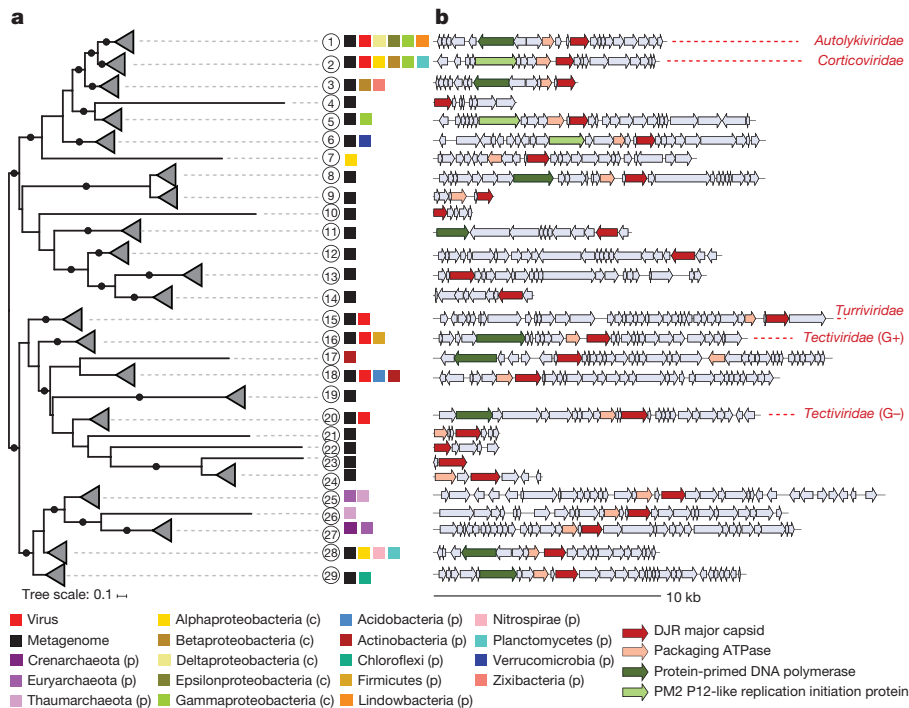


Figure 4 | DJR capsid viruses are far more diverse than the three currently recognized families, and include hosts in diverse bacterial and archaeal phyla. a, Phylogeny of 442 bacterial and archaeal DJR virus capsid proteins (sequences in Supplementary Data 1), including representatives of three previously described DJR virus families and sequences newly identified here; group numbers are assigned to each

of DJR virus capsids in genomes of nine additional phyla, including the two most abundant groups in the marine environment, the Alphaproteobacteria and the Thaumarchaeota (Fig. 4a and Methods). Moreover, analyses of marine metagenomes reveal that the environmental diversity of bacterial and archaeal DJR capsids exceeds that of our reference panel by several-fold (Fig. 4a). To specifically and conservatively evaluate the diversity of bacterial and archaeal DJR viruses, we selected only sequences with strong support for structural similarity to these groups for further analyses, omitting a large cluster of putative eukaryotic DJR viruses and sequences with no detectable similarity to known proteins (Extended Data Fig. 8). DJR genomic neighbourhoods encompass other viral proteins, and carriage of the protein-primed polymerase, which is associated with the presence of covalently bound terminal proteins, is common across deeply divergent lineages (Fig. 4b). Members of these groups would thus also be subject to the protease-dependent extraction bias (Fig. 3c).

The discovery of the autolykiviruses provides insight into the nature of the non-tailed viruses that dominate the global surface ocean, and suggests that dsDNA bacterial and archaeal DJR viruses have been systematically excluded from discovery. By providing genome-sequenced isolates and optimized approaches for targeted recovery of additional diverse representatives, we address a major challenge for metagenomic surveys—the paucity of viral reference genomes necessary for the interpretation of the uncharacterized majority of sequence diversity and function²⁹. The extensive sequence diversity that we find among bacterial and archaeal DJR elements suggests that additional, culture-based reference sequences will be required to assess their true environmental diversity. The distinctively broad host ranges of members of the *Autolykiviridae* and related DJR elements also suggest that, if such viruses are capable of packaging host DNA, they may have an even more important role in facilitating the observed gene transfers between highly divergent bacteria in microbial communities³⁰ than the highly specific tailed viruses. Finally, the recovery of the non-tailed autolykiviruses represents a first step in revealing extensive missed

branch for reference, coloured blocks indicate hosts, black circles on branches indicate approximate likelihood-ratio test (aLRT) branch support ≥ 0.9 . **b**, Element gene diagrams from each group show prophage host genome neighbourhoods and metagenome contigs often contain additional genes common to DJR elements (contig information in Extended Data Table 2). G+, Gram-positive; G–, Gram-negative.

diversity in one of the two major ancient lineages of dsDNA bacterial viruses and suggests that their ecological and evolutionary importance for microbial systems is far greater than is currently recognized.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 August 2016; accepted 28 December 2017.

Published online 24 January 2018.

- Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000).
- Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res.* **239**, 136–142 (2017).
- Krupovic, M., Prangishvili, D., Hendrix, R. W. & Bamford, D. H. Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. *Microbiol. Mol. Biol. Rev.* **75**, 610–635 (2011).
- Brum, J. R., Schenck, R. O. & Sullivan, M. B. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* **7**, 1738–1751 (2013).
- Benson, S. D., Bamford, J. K. H., Bamford, D. H. & Burnett, R. M. Does common architecture reveal a viral lineage spanning all three domains of life? *Mol. Cell* **16**, 673–685 (2004).
- Krupovic, M. & Bamford, D. H. Archaeal proviruses TKV4 and MVV extend the PRD1-adenovirus lineage to the phylum Euryarchaeota. *Virology* **375**, 292–300 (2008).
- Pietilä, M. K. et al. Structure of the archaeal head-tailed virus HSTV-1 completes the HK97 fold story. *Proc. Natl Acad. Sci. USA* **110**, 10604–10609 (2013).
- Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* **479–480**, 2–25 (2015).
- Wikoff, W. R. et al. Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**, 2129–2133 (2000).
- Krupovic, M. & Bamford, D. H. Virus evolution: how far does the double β -barrel viral lineage extend? *Nat. Rev. Microbiol.* **6**, 941–948 (2008).
- Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl Acad. Sci. USA* **114**, E2401–E2410 (2017).
- Iranzo, J., Krupovic, M. & Koonin, E. V. The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. *MBio* **7**, e00978–16 (2016).
- International Committee on Taxonomy of Viruses. ICTV Master Species List v.1.3 <https://talk.ictvonline.org/files/master-species-lists/m/msl/6776> (2016).

14. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
15. Espejo, R. T. & Canelo, E. S. Properties of bacteriophage PM2: a lipid-containing bacterial virus. *Virology* **34**, 738–747 (1968).
16. Wommack, K. E., Hill, R. T., Kessel, M., Russek-Cohen, E. & Colwell, R. R. Distribution of viruses in the Chesapeake Bay. *Appl. Environ. Microbiol.* **58**, 2965–2970 (1992).
17. Andrews-Pfannkoch, C., Fadrosch, D. W., Thorpe, J. & Williamson, S. J. Hydroxyapatite-mediated separation of double-stranded DNA, single-stranded DNA, and RNA genomes from natural viral assemblages. *Appl. Environ. Microbiol.* **76**, 5039–5045 (2010).
18. Steward, G. F. *et al.* Are we missing half of the viruses in the ocean? *ISME J.* **7**, 672–679 (2013).
19. Labonté, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J.* **7**, 2169–2177 (2013).
20. Roux, S. *et al.* Towards quantitative viromics for both double-stranded and single-stranded DNA viruses. *PeerJ* **4**, e2777 (2016).
21. Peralta, B. *et al.* Mechanism of membranous tunnelling nanotube formation in viral genome delivery. *PLoS Biol.* **11**, e1001667 (2013).
22. Sun, L. *et al.* Icosahedral bacteriophage Φ X174 forms a tail for DNA transport during infection. *Nature* **505**, 432–435 (2014).
23. Saren, A.-M. *et al.* A snapshot of viral evolution from genome analysis of the *Tectiviridae* family. *J. Mol. Biol.* **350**, 427–440 (2005).
24. Thurber, R. V., Haynes, M., Breitbart, M., Wegley, L. & Rohwer, F. Laboratory procedures to generate viral metagenomes. *Nat. Protoc.* **4**, 470–483 (2009).
25. Castro-Mejía, J. L. *et al.* Optimizing protocols for extraction of bacteriophages prior to metagenomic analyses of phage communities in the human gut. *Microbiome* **3**, 64 (2015).
26. D'Herelle, F. Studies upon Asiatic cholera. *Yale J. Biol. Med.* **1**, 195–219 (1929).
27. Krupović, M. & Bamford, D. H. Putative prophages related to lytic tailless marine dsDNA phage PM2 are widespread in the genomes of aquatic bacteria. *BMC Genomics* **8**, 236 (2007).
28. Xue, H. *et al.* Eco-evolutionary dynamics of episomes among ecologically cohesive bacterial populations. *MBio* **6**, e00552–e15 (2015).
29. Brum, J. R. & Sullivan, M. B. Rising to the challenge: accelerated pace of discovery transforms marine virology. *Nat. Rev. Microbiol.* **13**, 147–159 (2015).
30. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480**, 241–244 (2011).
31. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. King, P. Weigele, J. Daily and J. Chodera for comments and suggestions; T. Soni and members of the Polz laboratory for assistance with sampling; S. Labrie for guidance in viral genome extractions and sequencing library preparation, and C. Haase-Pettingell for assistance with density gradients; N. Watson for electron microscopy; and R. Ratzlaff for discussions and the suggestion of electron microscopy of virus plaques in agar overlay. This work was supported by grants from the National Science Foundation OCE 1435993 to M.F.P. and L.K., the NSF GRFP to F.A.H. and the WHOI Ocean Ventures Fund to K.M.K.

Author Contributions K.M.K., F.A.H., L.K. and M.F.P. designed the study and planned experiments and analyses. K.M.K., L.K. and M.F.P. wrote the paper with contributions from all authors. K.M.K. conducted field sampling, isolations and experimental characterizations of lytic viruses. J.Y. conducted the statistical analyses of the viral decay experiment and wrote the scripts to visualize the infection matrix as a phylogeny-anchored network, which was based on the host ribosomal protein tree generated by P.A. W.K.C. and L.K. performed the quantification of significance of host sharing. F.A.H. performed isolation and characterization of active *Vibrio* DJR prophages. Bacterial genome sequencing libraries were prepared by M.B.C., assembled by P.A., and curated and annotated by P.A. and J.E. The viral genome sequencing libraries were prepared by K.M.K. and R.S.S., assembled by J.M.B. and K.M.K., and annotated and curated by J.M.B., K.M.K., J.E., W.K.C. and L.K. The viral metagenome sequencing libraries were prepared by K.M.K., and assembled and curated by P.A. and L.K. The bioinformatic analyses of microbial genomes and metagenomes for DJR capsid elements were performed by L.K. and K.M.K., and the visualization of the DJR network was performed by D.V. M.B.C. provided field and laboratory technical support throughout. Although specific contributions are highlighted for each author, all authors contributed in additional ways through contributions to figures, analyses, discussion of results and comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to M.F.P. (mpolz@mit.edu) or L.K. (libusha.kelly@einstein.yu.edu).

Reviewer Information *Nature* thanks J. Fuhrman, E. V. Koonin and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Isolation, culturing and sequencing of bacteria and viruses. Bacteria and viruses were collected from the littoral marine zone at Canoe Cove, Nahant, Massachusetts, USA, on 22 August (ordinal day 222), 18 September (261) and 13 October (286) 2010.

Bacteria were collected using previously described size-fractionation and selective-medium cultivation-based methods³². Bacterial genome libraries were prepared for sequencing using a tagmentation-based approach and 1–2 ng input DNA per isolate, as previously described³³. Genomes were sequenced in multiplexed pools of 50–60 samples per Illumina HiSeq lane. Accession numbers for all bacterial genomes are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Bacterial phylogenetic relationships were determined by concatenation of ribosomal proteins and *hsp60* sequences. For all strains with available genome sequences (278), ribosomal proteins were extracted from genomes with *hmmsearch*³⁴ and aligned with MAFFT³⁵ as previously described³⁶. Full-length *hsp60* sequences were also extracted from these genomes using *hmmsearch* with default parameters and the *Cpn60* *hmm* (PF00118) from Pfam³⁷. The *hsp60* sequences were aligned using the *mafft-fftinsi* algorithm. Sanger-sequenced *hsp60* fragments from 40 strains that lacked genome sequences were added to this alignment using the *mafft-fftinsi* algorithm with the *-addfragments* option. The *hsp60* alignment was concatenated to the ribosomal protein alignment and used to create a phylogeny using RAxML under a partitioned general time reversible (GTR) model (options: *-q -m GTRGAMMAX*)³¹. Shimodaira–Hasegawa (SH)-like supports were calculated using RAxML and taxonomy was assigned by manual inspection³⁶.

Viruses were collected using a previously described iron flocculation approach³⁸, using 4-l sample volumes, 0.2- μ m pre-filtration to remove bacteria, 0.2- μ m filters for floc capture, and oxalate solution for resuspension to maintain virus viability. Isolation of viruses was performed by directly plating virus concentrates in agar overlays on hosts from each of the same days, as follows. Iron-oxalate concentrate volumes equivalent to 15 ml of seawater were mixed with 150 μ l of overnight host culture and 2 ml molten top agar to form host lawns in overlay and allow for plaque formation (top agar: 52 °C, 0.4% agar, 5% glycerol, in 2216 Marine Broth (2216MB); bottom agar: 1% agar, 5% glycerol, 125 ml l⁻¹ of chitin supplement (40 g l⁻¹ coarsely ground chitin, autoclaved, 0.2- μ m filtered) in 2216MB). After incubation for two weeks, all plaques from plates containing fewer than approximately 25 plaques were archived, and a random subsample of each distinct plaque morphotype was archived for plates with more plaques. Plaque plugs were first eluted in 200 μ l of 2216MB, a subsample of 150 μ l was then filtered to remove bacteria for storage of virions at 4 °C, and the remainder was supplemented with 50% glycerol for storage at –20 °C. For purification and amplification, one archived plaque was randomly selected from all available plaques for each host and serially re-passaged at least three times, with stocks preferentially recovered from –20 °C archives. In a small number of cases, multiple plaques were purified for a given host and these are identifiable by the nomenclature described below.

Sequencing and genome analysis of viruses was as follows. In brief, high-titre plate lysates of serially purified viruses were concentrated using 30-kDa centrifugal filter units (Millipore, Ultracel 30K, UFC903024) and washed with 1:100 2216MB to reduce salts for nuclease treatment. Concentrates were brought to approximately 500 μ l using 1:100-diluted 2216MB and then treated with DNase I and RNase A for 65 min at 37 °C to digest unencapsidated nucleic acids. Nuclease-treated viral lysates were extracted by addition of 1:10 final volume of SDS mix (0.25 M EDTA, 0.5 M Tris-HCl (pH 9.0), 2.5% sodium dodecyl sulphate), 30 min incubation at 65 °C; addition of 0.125 volumes 8 M potassium acetate, 60 min incubation on ice; addition of 0.5 volumes phenol–chloroform; and recovery of nucleic acids from aqueous phase by isopropanol and ethanol precipitation. Genomes were fragmented by sonication, libraries sequenced in multiplexed pools using Illumina MiSeq and HiSeq technologies, assembled using CLC Genomics Workbench v6.5.1 and v8.5.1 and CLC assembly cell v4.4.2.133896, and manually curated to standardize genome start positions for the *Caudovirales*. Bioinformatic analyses indicate that all sequenced viruses, except autolykiviruses, are members of the *Caudovirales*.

The viral-strain naming convention is described using the example of 1.008.O._10N.286.54.E5, with specific identifiers separated by a full stop. The first position (here '1') represents a unique identifier for each independent plaque isolated from a given host from the initial exposure of that host to an environmental virus concentrate. The second position (here '008') represents a unique working ID for a host strain. The third position (here 'O') indicates a unique sublineage generated from a single plaque during viral serial purification, for example, owing to the emergence of multiple plaque morphologies. Following the underscore is the full strain ID of the host of isolation. Viral genome accession numbers are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Characterization of virions of autolykiviruses. Morphology of the autolykiviruses was determined by thin-section electron microscopy (TEM) of a representative member, 1.008.O (Fig. 1a, b and Extended Data Fig. 1a–c). TEM was performed once on a single agar overlay. Viruses were visualized by generating plaques in the agar overlay and then fixing the overlay for 14 h by addition of fixative (2.5% glutaraldehyde, 3% paraformaldehyde with 5% sucrose in 0.1 M sodium cacodylate buffer (pH 7.4)). The top agar was collected into 0.1 M sodium cacodylate buffer (pH 7.4), pelleted and washed in sodium cacodylate buffer, soaked overnight in 1% OsO₄ in veronal-acetate buffer, stained en bloc overnight with 0.5% uranyl acetate in veronal-acetate buffer, dehydrated and embedded in Embed-812 resin. Ultrathin sections were prepared on a Leica Ultracut E microtome with a Diatome diamond knife, stained with 2% uranyl acetate and lead citrate. Sections were examined using an FEI Tecnai Spirit electron microscope at 80 kV and photographed with an AMT CCD camera. The selection of a region containing both uninfected and lysed cells allowed for capture of multiple stages of infection without the need for optimization of infection course timing. Capsid measurements were made using ImageJ³⁹. Ten virus particles in a single image (magnification of 98,000 \times) were each measured at three different cross sections and the average calculated for each virus was used to determine the overall mean and standard deviation. Observations of inner-core and tail-tube-like structures are consistent with those previously described for the lipid-bilayer-containing non-tailed corticovirus⁴⁰ and the non-tailed tectivirus PRD1²¹, respectively.

Ultracentrifugation in Optiprep iodixanol gradient medium (Sigma D1556) was used to determine the density of representative autolykiviruses. This medium was selected on the basis of previous demonstrations of sensitivity of some viruses, including the related corticovirus PM2, to caesium chloride and sucrose gradients²⁴. Use of iodixanol also allowed for direct culture-based assay of viral activity in density gradient fractions without prior dialysis, as required for other density gradient media. Density gradients were prepared using artificial seawater (ASW, Sigma S9883, 40 g l⁻¹) diluent, as follows: eight density layers spanning 20% to 54% iodixanol were manually laid in Seton 7030 tubes and loaded with 500 μ l of polyethylene glycol (PEG)-precipitated viral concentrate resuspended in ASW. Samples were centrifuged for 10 h at 20 °C and 35,000 r.p.m. in a SW41 swinging-bucket rotor in a Beckman L8M ultracentrifuge. Density gradient fractions were collected using the Biocomp Piston Gradient Fractionator (BioComp Instruments Inc.) and densities determined as mass per 100 μ l volume, using a standard laboratory pipette⁴¹. Densities for each virus were defined as the fraction with greatest plaque-forming activity in an agar-overlay assay of density fractions collected from a single density column. Data shown in Fig. 3b are from a single experiment that included five representative autolykiviruses (1.008.O, 1.18 g ml⁻¹; 1.020.O, 1.16 g ml⁻¹; 1.044.O, 1.19 g ml⁻¹; 1.080.O, 1.15 g ml⁻¹; and 1.249.A, 1.20 g ml⁻¹), processed together in a common centrifugation run. This test set also included the lipid-containing PRD1 tectivirus as an internal control, however, this tectivirus showed a bimodal distribution of peak infectivity (1.15 g ml⁻¹ and 1.21 g ml⁻¹) associated with distinct plaque morphologies, observations consistent with the presence of both the expected lower-density PRD1 and a contaminating higher-density tailed prophage in the stock. A subsequent independent iodixanol density gradient centrifugation experiment that included both an autolykivirus (1.044.O) and a tailed virus (1.255.O) also showed a lower density for the autolykivirus (1.17 g ml⁻¹) than for the tailed virus (1.22 g ml⁻¹).

The buoyant densities of corticovirus PM2⁴² and marine vesicles⁴³ in iodixanol (1.16 g ml⁻¹ and 1.15–1.19 g ml⁻¹, respectively) indicated in Fig. 3b are based on literature values, and the range for tailed viruses is inferred from the literature value for PM2 assuming a linear correspondence between the density of any virus in iodixanol and its density in caesium chloride. Using values from a study⁴² in which densities for PM2 were determined in both iodixanol (1.16 g ml⁻¹) and caesium chloride (1.28 g ml⁻¹), we infer that the tailed viruses targeted at the 1.35–1.50 g ml⁻¹ interface in caesium chloride would span densities from 1.22–1.36 g ml⁻¹ in iodixanol. We note that whereas there is extensive data showing that lipid-containing viruses are outside of the range commonly targeted for tailed viruses in caesium chloride, data on tailed virus densities measured in both caesium chloride and iodixanol are lacking. We therefore caution that our approximation is a guideline, and that future studies attempting targeted isolation from iodixanol density gradient media should first ensure that iodixanol also yields the desired separation to the same extent as caesium chloride.

Chloroform sensitivity of members of the *Autolykiviridae* was assessed using a 0.2-volume ratio of chloroform to virus-containing solution, as commonly applied to viral concentrates to eliminate bacterial contamination²⁴. The test set of viruses (Fig. 3a) included five representative autolykiviruses (1.008.O, 1.020.O, 1.044.O, 1.080.O, 1.249.A; all with genome size of approximately 10 kb) and four representative tailed viruses, including the *Escherichia coli* siphovirus Lambda, and three representative Vibrionaceae viruses: a siphovirus (1.082.O; approximately

36 kb), a myovirus (1.170.O; approximately 134 kb) and a podovirus (2.130.O; approximately 76 kb). All viruses were tested in three independent replicates, and each replicate included a pair of samples, with and without chloroform exposure. Chloroform-treatment samples were mixed with chloroform, all samples were gently vortexed for 6 s, incubated at room temperature, and mixed twice by finger-flick over 2 h. Samples were then centrifuged at 5,000 g for 5 min and the activity was assessed using a dilution series of drop spots (5 µl) on host agar-overlay lawns, including no-virus controls to allow detection of chloroform carry-over.

The effect of protease treatment on recovery of nucleic acids from tailed and autolykivirus viruses was assessed using a method commonly applied to generate marine viral metagenomes^{44,45}. The test set of viruses (Fig. 3c) included five representative autolykiviruses (1.008.O, 1.020.O, 1.044.O, 1.080.O, 1.249.A; all approximately 10 kb) and three representative tailed *Vibrionaceae* viruses: a siphovirus (1.082.O; around 36 kb), a myovirus (1.170.O; about 134 kb) and a podovirus (2.130.O; approximately 76 kb). Each viral concentrate was extracted three times independently, each time with samples in a different order, as follows. Lysates were nuclease-treated in 50 µl reactions containing 1 × Turbo DNase buffer, 1 µl Turbo DNase (Ambion AM2239), 0.5 µl RNase A (Thermo Scientific EN0531) and incubated at 37 °C for 60 min. Nuclease reactions were halted by addition of EDTA to 100 mM and heat inactivation for 10 min at 75 °C. Samples treated with protease received 0.5 µl of proteinase K (Epicentre MPRK092) per 62.5 µl reaction; all samples were incubated at 65 °C for 20 min. Four sub-aliquots for each virus sample and treatment were pooled to a final volume of 250 µl, mixed with 1 ml Wizard PCR Preps DNA Purification Resin (Promega A718A), loaded onto Wizard Minicolumns (Promega A721B), washed with 2 ml 80% isopropanol and collected by centrifugal wash with 80 °C TE buffer. dsDNA concentrations were quantified by fluorescence using the Quant-iT PicoGreen kit as per the manufacturers' protocol (ThermoFisher Scientific P7589). Products were visualized by agarose gel electrophoresis (0.7% agarose, 0.5 × TBE, 90 V for 90 min) with load volumes normalized to 50 ng across proteinase K-treated samples and all corresponding non-proteinase K-treated replicates loaded at equal volume to their treated replicates from the same independent experiment. The included gel image (Fig. 3c, gel source data are shown in Supplementary Fig. 1) is a representative experiment that contains a single replicate of each virus, for this representative experiment the Kruskal–Wallis test implementation of the Kruskal–Wallis test in R (v3.3.0) was used to test for differences between autolykiviruses and tailed viruses in the fold difference in DNA recovered with protease compared to without protease, measured by PicoGreen fluorescence (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 5$, d.f. = 1, $P = 0.02535$; $n = 5$ and 3, median = 8.35 and 0.97, for autolykiviruses and tailed viruses, respectively).

Evaluation of decay rates of autolykiviruses and tailed viruses. Decay of eight viruses (five autolykiviruses and three tailed) was monitored in ASW in borosilicate vials at room temperature in the dark over 34 days. Activity was measured by drop-spot plating of three independent serial dilutions of each of four replicate samples of each virus on days 0, 1, 10, 20 and 34. A linear mixed model for the decay data was fit using the lme4 package^{46,47} in R⁴⁸, with \log_{10} of the PFU counts as the response variable, an intercept (starting PFU) and slope over time (decay rate) as fixed effects, and intercept and slope for each virus as well as intercept and slope for each bottle nested in each virus as random effects. Decay rates measured in log loss per day (t) over the observation period were variable among viruses and substantially higher for two of the autolykiviruses (1.008.O, $-0.03t$; 1.020.O, $-0.08t$) than for the other autolykiviruses (1.044.O, $-0.02t$; 1.080.O, $-0.01t$; and 1.249.A, $-0.01t$) and the tailed viruses (podovirus 2.130.O, $-0.01t$; myovirus 1.170.O, 0t; siphovirus 1.082.O, 0t); 95% conditional predictive intervals of the autolykiviruses showed no overlap with the myovirus or siphovirus, nor did those for autolykiviruses 1.008.O and 1.020.O show overlap with the other autolykiviruses.

Annotation of DJR element genomes, contigs and genomic neighbourhoods. Open reading frames for all virus, plasmid, prophage and metagenomic contigs were identified using Prodigal⁴⁹ v2.6.1 with the $-p$ meta option. Elements not sequenced as part of this study were recovered as follows: parent nucleotide sequences of all DJR proteins with accession numbers were downloaded manually through NCBI Batch Entrez; where DJR proteins occurred in microbial genome backbones, regions of 20 kb centred around the DJR were downloaded. Proteins called *de novo* during this work from metagenomic contigs were re-associated with their parent contigs. Protein sequences derived from the OM-RGC were linked back to their metagenomic assemblies using the Tara Oceans companion website tsv table (ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq_release.tsv.gz)⁵⁰. Metagenomic assemblies were downloaded from the European Nucleotide Archive (<http://www.ebi.ac.uk/ena/about/tara-oceans-assemblies>)^{50,51} and scaffolds associated with each hit were extracted from the assemblies. Clusters of homologues were identified by performing an all-by-all BLASTp, requiring a minimum bitscore of 50, and clustering all pairs unweighted using Markov cluster

algorithm (MCL)⁵² v14.137 with an inflation parameter set to 1.4. Structural annotations were performed using the Phyre2⁵³ webportal and, for a subset of proteins, HHpred⁵⁴ through the MPI Bioinformatics Toolkit⁵⁵. Sequence similarity-based annotations were performed using BLASTp searches against NCBI RefSeq Virus genome, the NCBI Batch Web Conserved Domain⁵⁶ search tool, and EggNOG-Mapper⁵⁷. Sequences were also annotated with InterProScan⁵⁸ v5.17-56.0 using the *iplookup*, *goterms* and *pathways* options, and including two optional databases, TMHMM and SignalP⁵⁹, in addition to the 13 default databases. All annotations and cluster information are provided in Supplementary Table 1. Genome diagram figures were prepared using the GenoPlotR⁶⁰ package in R and refined in Adobe Illustrator.

Detailed analyses of the gene of unknown function adjacent to the gene encoding the protein-primed DNA polymerase (pDNApol) in autolykivirus genomes suggest that it encodes the terminal protein that is necessary for the protein-primed DNA polymerases to initiate replication. The observations that: (1) this is a core gene shared by all autolykiviruses; (2) the secondary structure predictions for the encoded protein are consistent with other terminal proteins; and (3) it is located adjacent to the pDNApol, as is the gene for the terminal protein in PRD1 and phi29, are strong evidence to suggest that this orthologous cluster in the autolykiviruses represents a novel terminal protein.

Construction of alignments and phylogenetic trees. To evaluate nucleotide diversity among members of the *Autolykiviridae*, we performed whole-genome alignments using the EMBL–EBI implementation of Clustal Omega^{61–63} and PhyML⁶⁴ with SMS⁶⁵ v1.8.1 (Extended Data Fig. 2). To evaluate the relationship of the autolykiviruses to known DJR viruses, and the support for their establishment as a new viral family, we evaluated gene trees for three conserved genes representative of the structural and replication functions of these viruses, the major capsid protein and packaging ATPase, and pDNApol, respectively. We included only bacteria- and archaea-infecting viruses, excluding eukaryote-infecting DJR viruses, and defined membership in each of the gene trees as described in ref. 12, with the exception that we excluded members of the *Caudovirales* from the pDNApol tree. When protein sequences were available in the pVOG database⁶⁶, these were used, otherwise sequences were downloaded from NCBI RefSeq¹⁴. Included in the major capsid protein tree were members of the *Tectiviridae*, *Corticoviridae* and *Turriviridae*; of these only the Gram-positive bacteria-infecting members of the *Tectiviridae* were included in a pVOG (VOG0339), with the others acquired from NCBI RefSeq. Included in the packaging ATPase tree were viruses from the *Tectiviridae*, *Corticoviridae* and *Turriviridae* families, and *Sphaerolipoviridae*; the pVOGs (VOG4814, VOG0337) for this gene did not include the *Corticoviridae* or *Turriviridae*, which were acquired from NCBI RefSeq. Included in the pDNApol tree were viruses in the *Tectiviridae* and *Ampullaviridae* families, and Salterprovirus (VOG0334). All sequences, including those of the *Autolykiviridae* virus representatives, were clustered using *usearch* (*-cluster* *_fast query.fasta* *-sort length* *-id 0.9* *-centroids* *nr.fasta* *-uc clusters.uc*) and representative members of each cluster were selected for consistency across gene trees where possible⁶⁷. All alignments and phylogenetic trees were constructed using the alignment, curation and maximum likelihood tree-building pipeline workflow referred to as eggNOG41 in the ETE v3.0.0b36 tree-building tool⁶⁸, implementing Clustal Omega⁶¹, Muscle⁶⁹, MAFFT v5⁷⁰, M-Coffee⁷¹, trimAl⁷² and PhyML 3.0⁶⁴, and executed as: *ete3 build -a my_sequences.fasta -w eggnog41 -o results/*⁷³.

Characterization of the host range of autolykiviruses. Host ranges of the autolykiviruses and tailed viruses were characterized using drop-spot assays, and a host panel that included all hosts of isolation of the purified viruses. Viruses were applied to agar-overlays of host lawns as triplicate randomized-position spots in 150-mm Petri dishes using 96-spot blotters (BelArt, Bel-blotter 96-tip replicator, 378760002). Activity was monitored for all spots on days 1, 2, 3, 7, 14, 21 and 30 by marking boundaries of clearings on the Petri dish. At the termination of the experiment, all positives were called, blind to corresponding replicates, and sizes of clearings at each time point were recorded. Potential for cross-contamination was assessed by visual inspection and considered in final conservative manual curation of 'positive' infection calls. As a result, some cases with 3/3 positive replicates were discarded due to high probability of cross-contamination and some cases with 2/3 positive replicates were included when, for example, these were the only positives on a test plate.

The infection dataset, which was curated as described above, including only viruses that infected their host of isolation again in the host range assay and derived from independent plaques in the original isolation, included 247 viruses (Fig. 2). For statistical comparisons of infections of autolykiviruses and tailed viruses, only the 241 sequenced viruses were included. Four autolykiviruses were excluded from infection analyses, because they either represent genomically identical sublineages of a member included in the analyses (1.107.A, 1.107.B and 1.249.B) or because they did not infect their original host of isolation in the large-scale host range assay (1.095.O).

The kruskal.test implementation of the Kruskal–Wallis test in R (v3.3.0) was used to test for differences between autolykiviruses and tailed viruses in number of hosts (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 38.9724$, d.f. = 1, $P = 4.298 \times 10^{-10}$; $n = 17$ and 224, median = 34 and 2, for *Autolykiviridae* and tailed viruses, respectively) and number of host species (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 94.9497$, d.f. = 1, $P < 2.2 \times 10^{-16}$; $n = 17$ and 224, median = 4 and 1, for autolykiviruses and tailed viruses, respectively); for the test of the number of host species, assignments were based on the species defined in Fig. 2. For comparisons of average genome identity of hosts infected by the autolykiviruses and the tailed viruses, only infections between fully sequenced bacteria and viruses with >1 host were included (two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 26.1429$, d.f. = 1, $P = 3.171 \times 10^{-7}$; $n = 16$ and 106, median = 93.04% and 99.97%, for autolykiviruses and tailed viruses, respectively). Evaluation of the time to detection of plaques in the host range assay also showed that, on average, the autolykiviruses required three times longer than tailed viruses to become detectable in culture ($n = 498$ infections by 17 autolykiviruses, $n = 844$ infections by 224 tailed viruses; median = 3 days and 1 day, for autolykiviruses and tailed viruses, respectively; two-sided Kruskal–Wallis rank-sum test; $\chi^2 = 374.7938$, d.f. = 1, $P < 2.2 \times 10^{-16}$, Fig. 3d and Extended Data Fig. 6).

In order to visualize the infection network with reference to the host phylogeny, we used iTOL⁷⁴ to generate an inverted circular representation of the host phylogeny and combined this with a Gephi-based ordered infection network representation generated using custom scripts in R and the packages igraph and rgraphviz^{75,76}. In Gephi, all nodes were connected to a dummy hub node at the centre of the network, host nodes were fixed to the periphery and ordered to match the iTOL tree, and virus nodes were connected to the hosts that they were able to infect. The Force Atlas 2 layout was used to adjust the position of the virus nodes in the network.

Quantification of the significance of host sharing. For each pair of viruses, X and Y, that share at least one host, the significance of the overlap in host range was calculated as follows. Assuming that Y infects K hosts out of a population of N hosts, and X infects n randomly selected hosts, the probability that X and Y will coinfect k or more hosts is given by $P = f(k; N, K, n)$, where f is the probability mass function of the hypergeometric distribution. We set k , N , K and n to their empirically observed values and take the negative log of P as the significance of coinfection between X and Y.

Characterization of active DJR prophages in *Vibrio*. DJR prophages were isolated and sequenced from *V. kanaloae* (5S-149; contig_10: 28913–43245) and *V. cyclitrophicus* (10N.286.55.C7; contig_73: 31709–46046) as follows: 1 ml of overnight host culture grown in 2216MB was inoculated into a 2-l baffled flask containing 1 l of fresh 2216MB. Cultures were grown with shaking at room temperature for seven days to allow for natural induction. Cells were removed using centrifugation (spun in sterilized 1-l polypropylene canisters at 5,000g for 15 min at 20°C using a JLA-8.1000 rotor in a Beckman Coulter Avanti J-20 XP centrifuge) followed by filtration of the supernatant through a 0.2- μ m vacuum filter (Corning 1,000 ml sterile Vacuum Filter/Storage Bottle System, 0.22- μ m PES Membrane). Cell-free 0.2- μ m filtrate was concentrated using PEG precipitation, as follows: 10% w/v of PEG 8000 (Sigma–Aldrich) was added to 700 ml of the filtrate at 0.6 M NaCl, solution was incubated with shaking at room temperature until PEG was visibly dissolved (3 h), incubated overnight at 4°C, after which the solution was centrifuged at 8,000g for just under 4 h at 20°C. The pellet was then collected with a sterile transfer pipette, resuspended in a final volume of 4 ml 0.02- μ m-filtered ASW (ASW, 40 g l⁻¹ Sigma Sea Salt solution prepared in sterile water) and stored at 4°C. A total of 0.7 ml of the PEG-concentrated sample was purified using iodixanol-based density ultracentrifugation (density gradient 20–54% iodixanol (OptiPrep) in ASW, centrifuged in a Beckman L8M centrifuge in an SW41 rotor for 10 h at 20°C at 35,000 r.p.m.). Gradients were unloaded as 26 fractions using a Biocomp Piston Gradient Fractionator (BioComp Instruments). Densities for each fraction were determined as mass per volume using a standard laboratory pipette⁴¹. Aliquots of each fraction were DNase-treated in 50- μ l reaction volumes with 1 \times TURBO DNase buffer and 1 μ l TURBO DNase and incubated at 25°C overnight, followed by addition of fresh TURBO DNase (1 μ l) and further incubation at 25°C for 2.5 h. DNase treatment was validated using gel electrophoresis of treated and untreated genomic DNA controls in comparable iodixanol solutions. DNA extractions were carried out as follows: 0.02- μ m-filtered ASW was added to reach a final volume of 100 μ l; nuclease activity was halted by addition of 1/10 final volume of hot SDS mix, incubated at 75°C for 10 min, then at 65°C for 20 min; proteins were degraded by addition of 1 μ l proteinase K per 100 μ l of reaction volume and incubated at 65°C for 20 min; DNA was recovered by addition of 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) with standard ethanol washes and elution in 20 μ l 0.2- μ m-filtered Elution Buffer (EB, Qiagen). Density fractions for sequencing were selected on the basis of a PCR assay using major capsid protein-specific primers for each element: extracted DNA from

fraction 12 (density = 1.19 for 5S-149 and 1.18 for 10N.286.55.C7) exhibited the brightest PCR band, suggesting the highest prophage concentration.

Final DNA extraction concentrations were quantified using a NanoDrop (5S-149 fraction 12 = 92.5 ng μ l⁻¹ and 10N.286.55.C7 fraction 12 = 24.9 ng μ l⁻¹). Major capsid gene-specific primers were ordered from IDT, with sequences as follows:

5S-149_MCP_F2, 5'-ACAGTTCACACAAGCGGGTC-3'; 5S-149_MCP_R2, 5'-AGTTCGCTGTGATAACGCCTA-3'; 10N.286.55.C7_MCP_F2, 5'-TCTTTTACGGGGACGGGCTA-3', 10N.286.55.C7_MCP_R2, 5'-CGCATATCTTC AAGCGCACG-3'.

Sample libraries were prepared for sequencing using the same tagmentation-based approach used for the bacterial genomes and ultimately multiplexed along with bacterial genomes on a single Illumina HiSeq lane. Sequenced reads were quality trimmed and mapped back to the reference genome of each lysogen to identify the full prophage region using CLC Genomics Workbench v8.5.1. *De novo* assemblies of reads also assembled the entire prophage into a single contig, which revealed the circular topology of the excised elements.

Metagenome preparation. An environmental sample was collected for metagenome preparation on 26 October 2014 (ordinal day 299) at Nahant, Massachusetts, USA. Eight replicate 4-l samples were collected and pre-filtered using 0.2- μ m Sterivex filters; the filtrate was iron-chloride flocculated, collected on 0.2- μ m Isopore polycarbonate filters (Millipore, GTTP09030) and resuspended in 4 ml oxalate solution, as described in ref. 38. For metagenome preparation, 1-ml subsamples from each of the eight replicates were pooled and PEG-concentrated (mixed: 8 ml pooled replicate subsamples, 0.8 g PEG, 8 ml 0.02- μ m-filtered 1 M NaCl; dissolved at room temperature for 2.75 h; incubated overnight at 4°C; centrifuged at 8,000g for 40 min at room temperature; the supernatant was then removed and the pellet resuspended in 600- μ l 0.02- μ m-filtered ASW; incubated at 4°C); the sample contained abundant white precipitate. Virus activity in pre- and post-concentration samples was compared using agar-overlay plating and plaque counts with the indicator host 10N.261.45.B10 to assess potential losses due to precipitation and recoveries were found to be 79% ($n = 3$). Nuclease activity was confirmed in samples diluted 1:1 with 0.02- μ m ASW.

A metagenome (14N.299.NahantUnfrac) was prepared from the concentrated sample as follows. To remove unencapsidated nucleic acids, the concentrated sample was pelleted to remove precipitates, a 100- μ l subsample was removed and diluted 1:1 with 0.02- μ m-filtered ASW, supplemented with 2 μ l Turbo DNase and 2 μ l RNase and incubated for 45 min at room temperature, pelleted to remove additional precipitates, and supplemented with an additional 2 μ l Turbo DNase and incubated for an additional 85 min. Next, to inactivate nucleases, the sample was supplemented with 0.5 M EDTA to a final concentration of 15 mM EDTA and incubated at 75°C for 20 min. The sample was then extracted using the MasterPure DNA extraction kit (Epicentre MPRK092) with proteinase K following the manufacturers' recommended protocol, with the exception of including an extended overnight ethanol precipitation. PicoGreen quantitation showed a final concentration of 75.1 ng μ l⁻¹ in 20 μ l, representing an original volume of 1,333 ml of seawater. Sequencing libraries were prepared using the Nextera Tagmentation approach as previously described³³, with an input concentration of 2 ng. Libraries were sequenced on a full NextSeq lane with 76 by 76 paired-end reads, at the MIT BioMicro Center.

A low buoyant density metagenome (14N.299.NahantLF) was prepared from the pooled replicates by density fractionating the PEG-concentrated virus sample and pooling subsamples of three low buoyant density fractions for extraction, as follows. First, 350 μ l of PEG-concentrated viruses (equivalent to 4,666 ml of original seawater) was loaded onto an iodixanol (OptiPrep) density step-gradient (20–54% iodixanol in ASW buffer), and centrifuged in a Beckman L8M centrifuge with an SW41 rotor for 10 h at 20°C at 35,000 r.p.m. (this procedure yielded precipitates upon addition of the sample to the density gradient). Then, gradients were unloaded as 26 fractions using a Biocomp Piston Gradient Fractionator (BioComp Instruments Inc.). Densities for each fraction were determined as mass per volume using a standard laboratory pipette⁴¹. Density fractions for metagenome preparation were conservatively selected on the basis of activity on a host infected by most autolykiviruses in the collection, 10N.261.45.B10 (fractions 9, 10, 11, with densities of 1.15, 1.16, 1.17 g ml⁻¹, respectively), these size fractions were conservatively selected and are known to exclude some members of the *Autolykiviridae* (Fig. 3b) as well as members of the excising DJR prophages of *Vibrio* (see Characterization of active DJR prophages in *Vibrio*). Selected iodixanol fractions were pooled (975.8 μ l), nuclease-treated (1 \times TURBO DNase buffer, 2 μ l TURBO DNase per 100 μ l final volume, 1 μ l RNase A per 100 μ l final volume), incubated at 25°C for 3.25 h in 50- μ l reaction volumes, after which the nuclease activity was halted by addition of 1/10 final volume of hot SDS mix and incubation at 75°C for 10 min, 65°C for 20 min. The sample was then treated with proteinase K with addition of 1 μ l

per 100 µl of reaction volume, incubated at 65 °C for 20 min and the DNA was recovered by addition of 0.5 volumes of Agencourt AMPure XP beads (Beckman Coulter) with standard ethanol washes and elution in 20 µl PCR-grade water. The 14N.299.NahantLF extract contained 8 ng µl⁻¹ DNA as determined by fluorescence. Sequencing libraries were prepared as previously described³³, using 12 replicate reactions that each had 1.13 ng input DNA, with the following modifications: input DNA extract was enriched for larger fragments with a 0.6× bead-based size selection, extension time in the second PCR in the protocol was increased to 60 s, and bead-based size selection was used to enrich for ~615-bp-length fragments following pooling of all 12 reactions. Libraries were sequenced on a full Illumina MiSeq lane with 250 × 250 paired-end reads, at the MIT MicroBio Center.

Reads for both the 14N.299.NahantUnfrac and the 14N.299.NahantLF were prepared as follows. Quality-trimmed paired and unpaired reads were assembled using the `clc_assembler` command (v4.4.2.133896) in the CLC Assembly Cell (CLC bio) with default parameters. Open reading frames were called with Prodigal v2.6.1 using the `-p` meta flag and otherwise default parameters. This protocol yielded 239,907 and 642,418 total genes for the 14N.299.NahantUnfrac and 14N.299.NahantLF metagenomes, respectively. Accession numbers for both metagenomes associated with this study are provided in Supplementary Data 3 and are included under NCBI BioProject PRJNA328102.

Identifying additional diverse bacterial and archaeal virus DJR capsid sequences. In order to evaluate DJR viruses in metagenomes, we first generated a reference panel of diverse bacterial and archaeal DJR virus capsid sequences that could then be used in metagenomic searches. To achieve this, we combined manual and iterative hidden Markov model (HMM)-based sequence searches of public databases, with structural and phylogenetic analyses of 'hit' sequences to generate a high-confidence, extensively curated and diverse bacterial and archaeal DJR virus capsid reference sequence set.

Our searches were initialized with a seed set of 24 DJR reference sequences, including four autolykiviruses, one corticovirus, ten corticovirus-like putative prophages²⁷, one Gram-negative-infecting tectiviruses, three Gram-positive-infecting tectiviruses, two turriviruses, the two excising *Vibrio* prophages described here, and one *Vibrio* plasmid identified here as a DJR element. Jackhammer⁷⁷ (<https://www.ebi.ac.uk/Tools/hmmer/search/jackhammer>) searches against UniProt⁷⁸ were used to generate HMMs for further searches, as well as to identify additional diverse DJR candidates, as revealed in the taxonomy view. We manually curated each round of HMM building and stopped the iterative search before eukaryotic viral proteins, primarily phycodnaviruses, were included in the HMM. This step was taken to skew our search towards bacterial and archaeal representatives of the DJR capsids. A subset of 12 HMMs was next used to search against NCBI bacterial (21,476) and archaeal (772) genomes (GenBank⁷⁹ Genomes, May 2017, <ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/>). These HMMs included: diverse representatives of the seed set (one each of the *Autolykiviridae*, the *Corticoviridae*, corticovirus-like putative prophages, Gram-positive-infecting viruses of the *Tectiviridae*, Gram-negative-infecting viruses of the *Tectiviridae* and the *Turriviridae*), additional recovered sequences confirmed to be virus capsid-like DJRs by curation with Phyre2⁵³ and the MPI Bioinformatics Toolkit⁵⁵ implementation of HHpred⁵⁴ (one each from genomes of *Magnetospirillum*, *Opitutaceae*, *Sulfobacillus*, *Nitrososphaera* and *Alcanivorax*), and one eukaryotic *Chlorella* virus DJR. Protein sequences for all downloaded microbial genomes were generated using Prodigal with the `-p` meta flag and otherwise default parameters and searches performed using the `hmmsearch`³⁴ tool (hmmer v3.1b2). These searches yielded 818 combined total unique hits, which were reduced to 196 by automatic screening to first require: (1) a size of 200–400 amino acid residues, the expected bacterial/viral DJR capsid size; (2) no hits to repeat domains. Next, manual trimming was applied to remove proteins with other functional domain annotations and the remaining sequences were then curated for a DJR-capsid-like structure, as described above.

This starting dataset enabled us to identify additional sequences in groups of particular interest, such as the alphaproteobacteria and other bacterial viruses, using manual `blastp`⁸⁰ searches against the GenBank non-redundant protein database⁷⁹. All additional hits identified manually were curated using Phyre2⁵³ and HHpred^{54,55} to identify sequences related to DJR protein structures from the Protein Data Bank (PDB)⁸¹. The sequences that were retrieved represent diverse phyla of archaea and bacteria, including Euryarchaeota, Crenarchaeota, Thaumarchaeota, Proteobacteria (alpha, beta, delta, epsilon and gamma representatives), Acidobacteria, Actinobacteria, Chloroflexi, Firmicutes, Lindowbacteria, Nitrospirae, Planctomycetes, Verrucomicrobia and Zixibacteria. One additional phage sequence was also identified from an unpublished *Rhodococcus* phage, and was described as a tectivirus although identified as a siphovirus by the NCBI taxonomy identifier. These putative DJR capsid proteins, plus the seed set of DJR bacterial and phage capsid proteins, a total of 179 unique sequences (Supplementary Table 2; marked as 'Reference' in Extended Data Fig. 8a), comprise our reference set and were next used to search ten bacterial and viral metagenomes.

Identifying potential DJR capsid proteins in metagenomes. Using our expanded reference set of sequences, we took a two-pronged approach to identify DJR capsid proteins in metagenomes. All proteins in each of ten metagenomes representing marine bacterial and viral fractions from environmental samples were analysed as follows (Extended Data Table 1). First, we ran jackhammer (hmmer v3.1b2)³⁴ with default parameters for each sequence in each metagenome and extracted hits with a full-sequence score >20. This analysis identifies sequences that are closely related to each of our individual DJR proteins. Second, we built a HMM out of the DJR reference protein sequence alignment using `hmmbuild` and then used `hmmsearch` to screen all proteins in each metagenome iteratively for five iterations and extracted hits with a full-sequence score >20. This second analysis potentially identifies more distantly related DJR sequences. These approaches together yielded 43,734 total potential DJR sequences.

Identifying relationships between potential DJR capsid proteins. We next wanted to identify clusters of proteins that might represent novel environmentally relevant groups of DJR capsid-containing elements associated with bacteria and archaea. We therefore combined a series of annotation and curation approaches to focus on proteins with strong support for associations with either bacterial or archaeal hosts.

First, we screened all sequences using the NCBI Batch Web Conserved Domain⁵⁶ search tool (<https://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi>) with default parameters^{56,82}. We retained only metagenomic sequences with either no hits to any conserved domains, or hits to known DJR capsid superfamilies, specifically: Capsid_N (cl25189), Capsid_NCLDV (cl04526) and Phage_Capsid_P3 (cl20087). Next, we used `psiblast`⁸⁰ to compare each sequence to the PDB⁸¹ protein structure database with an *e* value cut-off of 1×10^{-4} and retained only metagenomic sequences that either had no hits to any structures, or hits to known DJR virus capsids. The DJR PDB IDs used were: 1hx6, 2vvf, 1m3y, 2bbd, 5j7o, 3sam, 3j31, 1cjd, 4il7, 1m4x and 1j5q. Together, these screens narrowed our set to 25,874 potential DJR sequences.

To increase confidence, we next clustered all proteins and curated these clusters on the basis of both confirmed structural similarity to DJR capsids and sequence similarity to known viruses, as follows. First, we performed an all-by-all BLASTp search with a bitscore cut-off of 50 or better and clustered all proteins using MCL with unweighted BLAST matches and inflation value of 1.5 (Supplementary Table 2). We next annotated all proteins by whether they could be identified as a DJR through the Conserved Domain search, `psiblast`, Phyre2⁵³ or HHpred^{54,55} (only performed for a small subset of sequences). We then screened the around 26,000 sequences against the NCBI RefSeq Viral database (<https://www.ncbi.nlm.nih.gov/genome/viruses/>) and annotated all sequences with a best bitscore ≥ 50 to a *Caudovirales* virus sequence as spurious. Combining these annotations, we identified all clusters for which the number of sequences annotated as DJR was greater than those annotated as spurious and retained only these clusters for additional curation. Next, all retained clusters were evaluated for evidence of false positives as identified by Phyre2⁵³ structural similarity searches, with a requirement for length >250 amino acids, 95% confidence identification, and 75% alignment coverage, and any clusters with >5% of sequences with false positives were discarded. These additional curations together yielded a total of 14,666 passing proteins, which were retained for network visualization (Extended Data Fig. 8) along with two additional protein sequences that were among our references and structurally confirmed as DJR sequences (GenBank accessions: AOI82551.1 and WP_060243308.1) but were captured in an MCL cluster that was discarded due to abundant hits to sequences with *Caudovirales* virus taxonomy identifiers. Notably, although these sequences had very high confidence assignments to DJR major capsid proteins by both Phyre2 (PDB hits for both sequences to corticovirus PM2 capsid 2w0c; 100% confidence, 25–26% identity, 93% alignment coverage) and HHpred (PDB hits for both sequences to corticovirus PM2 major capsid protein 2vvf, 100% probability, *e* < 2×10^{-47} , target coverage 97%), they both had BLASTp bitscores of 45.8 against large proteins in tailed cyanophage (GenBank accessions: YP_007675165.1 and YP_009325074.1).

To ensure that proteins selected for subsequent phylogenetic analyses (Fig. 4) were strongly supported as being associated with viruses of bacterial and archaeal hosts, we next evaluated protein clusters on the basis of similarity to known DJR sequences and structures. All DJR hits identified by Conserved Domain search, `psiblast`, Phyre2 or HHpred were classified as either eukaryotic or bacterial and archaeal. Clusters with structurally annotatable sequences were dominated by either bacterial- and archaeal- or eukaryotic-associated virus DJR assignments, therefore, if the sum of bacterial- and archaeal-associated DJR hits (PDB identifiers: 1cjd, 1gw7, 1gw8, 1hb5, 1hb7, 1hb9, 1hqn, 1hx6, 1w8x, 2bbd, 2vvf, 2w0c and 3j31; Conserved Domain identifier: Phage_Capsid_P3 superfamily) was greater than the number of eukaryotic- or virophage-associated hits (PDB identifiers: 1j5q, 1m3y, 1m4x, 3j26, 3kk5, 3sam, 4il7 and 5j7o; Conserved Domain identifier: Capsid_N superfamily, Capsid_NCLDV superfamily) then the cluster was classified as

bacterial- and archaeal-associated (42 clusters; 788 proteins); if the reverse was true the cluster was classified as eukaryotic-associated (32 clusters; 12,998 proteins); and if there were no identified matches to any DJRs, the cluster was classified as unknown (474 clusters, including singletons; 882 proteins). The network diagram (Extended Data Fig. 8) was generated using the Python package NetworkX and visualized using Gephi v0.9.1. The network structure was generated using the ForceAtlas 2 force directed layout method, with the option to prevent node overlap.

All clusters identified as bacterial- and archaeal-associated were then further curated to identify clusters for which, despite inclusion of some members with hits to bacterial and archaeal virus DJRs, there was a prevalence of sequences with no matches to DJR structures despite both Phyre2 and HHpred annotation. Finally, all proteins passing these filters were required to be unique and at least 200 amino acids in length for inclusion in the final DJR major capsid tree; this yielded a final set of 442 proteins (Fig. 4 and Supplementary Data 2, with ten procedural duplicate sequences identified in 'Notes' column).

To build the phylogenetic tree from these 442 sequences, we executed the ETE toolkit⁶⁸ eggNOG41 phylogenetic workflow, as described above. The eggNOG41 gene tree workflow is used to construct trees in the EggNOG orthology database⁷³ and is therefore appropriate to construct a tree for related but very diverse sequences, as we have with our DJR protein set. In brief, this workflow incorporates comparison of several multiple alignment tools, an alignment trimming step that removes columns with >10% gaps, and protein model selection before constructing a tree in PhyML. In PhyML, the workflow optimizes the topology, the branch lengths and rate parameters (transition/transversion ratio, proportion of invariant sites, gamma distribution parameter). Equilibrium amino-acid frequencies are estimated using frequencies defined by the substitution model (in this case, the JTT model), four substitution rate categories and aLRT branch supports are used to construct the final tree. The tree was visualized in iTOL, collapsed on the basis of average internal branch length of 2.0, and exported for figure preparation in Adobe Illustrator. To provide an overview of the genomic neighbourhoods of the putative DJR capsid proteins in our phylogenetic tree, we identified a representative virus, genome-neighbourhood or metagenomic contig for each of the 29 major branches or clades (Fig. 4b and Extended Data Table 2) and annotated these (Fig. 4b, Supplementary Data 1) as described above for the autolykiviruses.

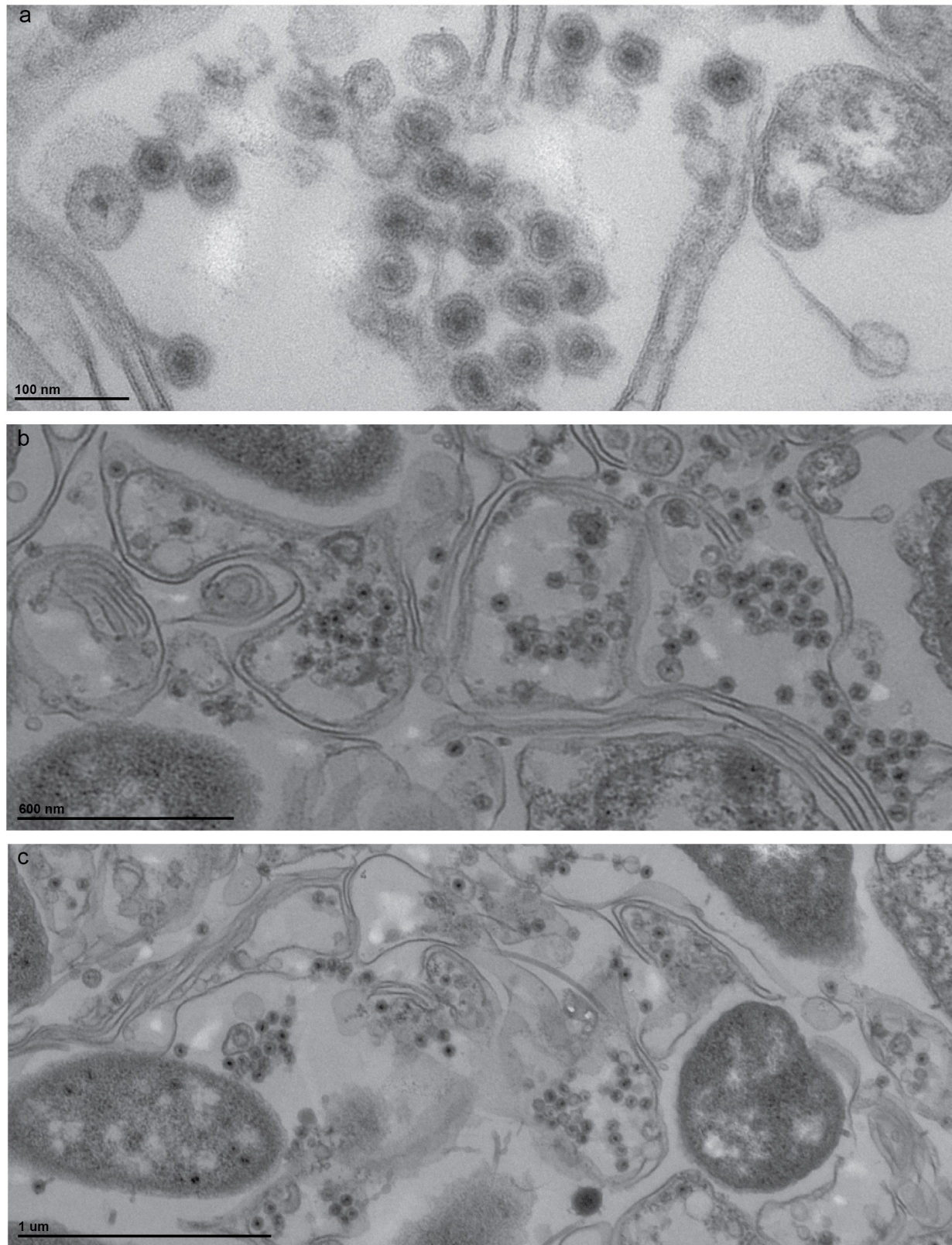
Code availability. All custom codes associated with this work are available from the authors upon request.

Data availability. Annotation information for the autolykiviruses and elements shown with genome diagrams is provided in Supplementary Data 1. Accession numbers, taxonomy and annotation of DJR capsid proteins included in the trees and network are provided in Supplementary Data 2. GenBank accession numbers for newly obtained sequences and previously published genomes included in Fig. 2 and Extended Data Fig. 6 are provided in Supplementary Data 3, with all new sequences associated with this work included under the Nahant Collection of NCBI BioProject with accession number PRJNA328102. Metagenomes used in this study are listed with citations in Extended Data Table 1 and include: *Tara* Oceans, viromes, ftp://ftp.imicrbe.us/projects/197/TOV_43_all_contigs_predicted_proteins.faa.gz; *Tara* Oceans, ocean microbiome reference gene catalogue, ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz; methane seep sediment, BioProject accession PRJNA290197; Rifle sediment, BioProject accession PRJNA288027; Mediterranean Sea virome, GenBank accessions, AP013358–AP014505; Mediterranean Sea metagenome, GenBank accessions, GU942957:GU943153; Chesapeake Bay virome, Sequence Read Archive accession, SRR4293227; NCBI environmental metagenomes, ftp://ftp.ncbi.nlm.nih.gov/blat/db/env_nr*.tar.gz; and two metagenomes generated in this study, Nahant light fraction viral metagenome, (deposited at GenBank under accession PDMW000000000; the version described here is PDMW010000000); and Nahant Viral Metagenome (deposited at GenBank under accession PDMX000000000, the version described here is PDMX010000000). All other data are available from the authors upon reasonable request.

32. Hunt, D. E. *et al.* Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081–1085 (2008).
33. Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS ONE* **10**, e0128036 (2015).
34. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
35. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
36. Hehemann, J.-H. *et al.* Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nat. Commun.* **7**, 12860 (2016).
37. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

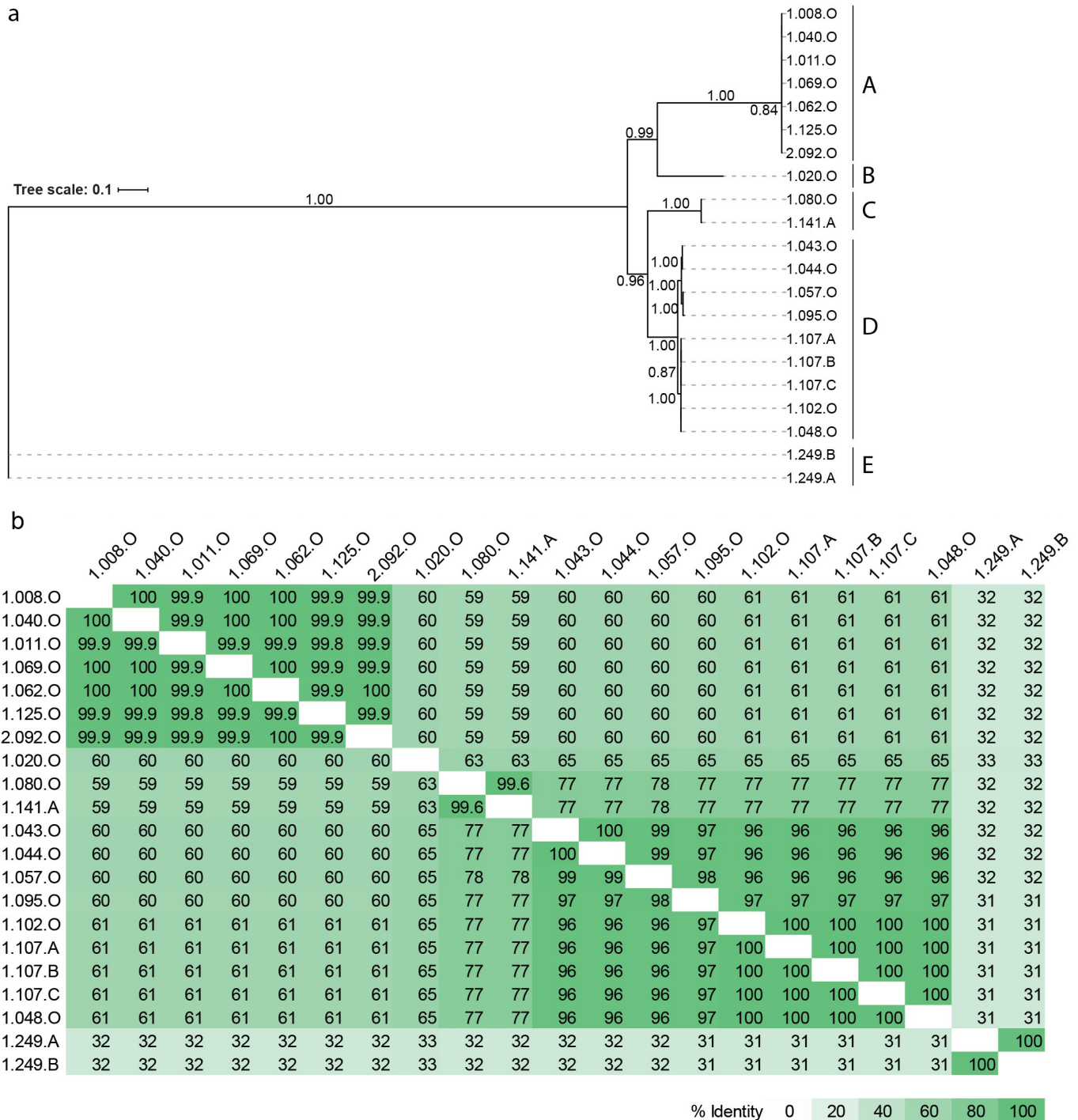
38. John, S. G. *et al.* A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.* **3**, 195–202 (2011).
39. Rasband, W. S. *ImageJ* (U.S. National Institutes of Health, 1997).
40. Silbert, J. A., Salditt, M. & Franklin, R. M. Structure and synthesis of a lipid-containing bacteriophage. 3. Purification of bacteriophage PM2 and some structural studies on the virion. *Virology* **39**, 666–681 (1969).
41. Lawrence, J. E. & Steward, G. F. in *Manual of Aquatic Viral Ecology* (eds Wilhelm, S. W., Weinbauer, M. G. & Suttle, C. A.) 166–181 (ASLO, 2010).
42. Kivelä, H. M., Männistö, R. H., Kalkkinen, N. & Bamford, D. H. Purification and protein composition of PM2, the first lipid-containing bacterial virus to be isolated. *Virology* **262**, 364–374 (1999).
43. Biller, S. J. *et al.* Bacterial vesicles in marine ecosystems. *Science* **343**, 183–186 (2014).
44. Hurwitz, B. L., Deng, L., Poulos, B. T. & Sullivan, M. B. Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environ. Microbiol.* **15**, 1428–1440 (2013).
45. Henn, M. R. *et al.* Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* **5**, e9083 (2010).
46. Bates, D. M. lme4: Mixed-effects modeling with R. <http://lme4.0.r-forge.r-project.org/IMMWR/lrgprt.pdf> (2010).
47. Bates, D. & Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
48. R Core Team. R: A Language and Environment for Statistical Computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, Vienna, Austria, 2016).
49. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
50. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
51. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**, 465–470 (2016).
52. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
53. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).
54. Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. Fast and accurate automatic structure prediction with HHpred. *Proteins* **77**, 128–132 (2009).
55. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44**, W410–W415 (2016).
56. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
57. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* **34**, 2115–2122 (2017).
58. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
59. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).
60. Guy, L., Kultima, J. R. & Andersson, S. G. E. genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334–2335 (2010).
61. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
62. Li, W. *et al.* The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res.* **43**, W580–W584 (2015).
63. McWilliam, H. *et al.* Analysis tool web services from the EMBL-EBI. *Nucleic Acids Res.* **41**, W597–W600 (2013).
64. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
65. Lefort, V., Longueville, J.-E. & Gascuel, O. SMS: smart model selection in PhyML. *Mol. Biol. Evol.* **34**, 2422–2424 (2017).
66. Graziotin, A. L., Koonin, E. V. & Kristensen, D. M. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. *Nucleic Acids Res.* **45**, D491–D498 (2017).
67. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
68. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
69. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
70. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005).
71. Wallace, I. M., O'Sullivan, O., Higgins, D. G. & Notredame, C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006).
72. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
73. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44** (D1), D286–D293 (2016).

74. Letunic, I. & Bork, P. Interactive tree of life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127–128 (2007).
75. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJ. Complex Systems*, 1695 (2006).
76. Vega Yon, G., Fabrega Lacoa, J. & Kunst, J. B. rgexf: Build, Import, and Export GEXF Graph Files. <https://cran.r-project.org/web/packages/rgexf/index.html> (2015).
77. Finn, R. D. *et al.* HMMER web server: 2015 update. *Nucleic Acids Res.* **43**, W30–W38 (2015).
78. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169 (2017).
79. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2013).
80. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
81. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
82. Marchler-Bauer, A. *et al.* CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225–D229 (2011).
83. Adriaenssens, E. & Brister, J. R. How to name and classify your phage: an informal guide. *Viruses* **9**, 70 (2017).
84. Clerissi, C. *et al.* Unveiling of the diversity of prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. *Appl. Environ. Microbiol.* **80**, 3150–3160 (2014).
85. Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, 1261498 (2015).
86. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358 (2017).
87. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**, 13219 (2016).
88. Anantharaman, K. *et al.* Analysis of five complete genome sequences for members of the class Peribacteria in the recently recognized Peregrinibacteria bacterial phylum. *PeerJ* **4**, e1607 (2016).
89. Mizuno, C. M., Rodriguez-Valera, F., Kimes, N. E. & Ghai, R. Expanding the marine virosphere using metagenomics. *PLoS Genet.* **9**, e1003987 (2013).
90. Ghai, R. *et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* **4**, 1154–1166 (2010).
91. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **44**, D67–D72 (2016).



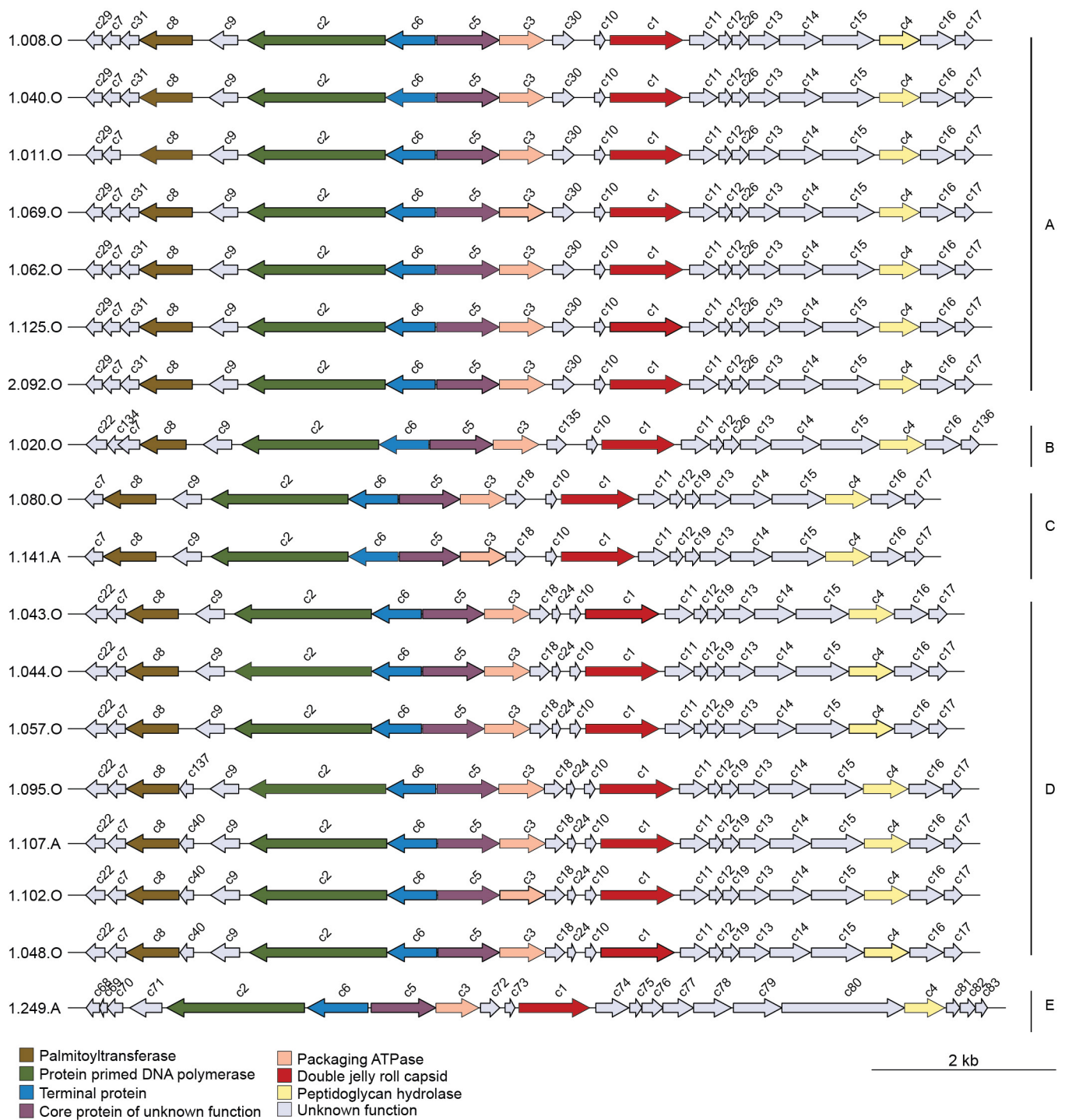
Extended Data Figure 1 | Members of the *Autolykiviridae* are non-tailed viruses that may form tail tubes on contact with cells. Thin-section electron microscopy of an agar overlay containing plaques of representative *Autolykiviridae* virus 1.008.O (see Methods for experimental details). **a**, Virus particles in contact with cell membranes are observed to occasionally possess tail-tube-like structures, whereas those

not in contact with cells do not. **b**, Lower magnification of same field of view as **a** shows that tail-tube-free virions are more common than those with tail tubes. **c**, Lower magnification view of virion in Fig. 1b also shows that the presence of the tail tube is associated with cell contact and is not observed in nearby virions.



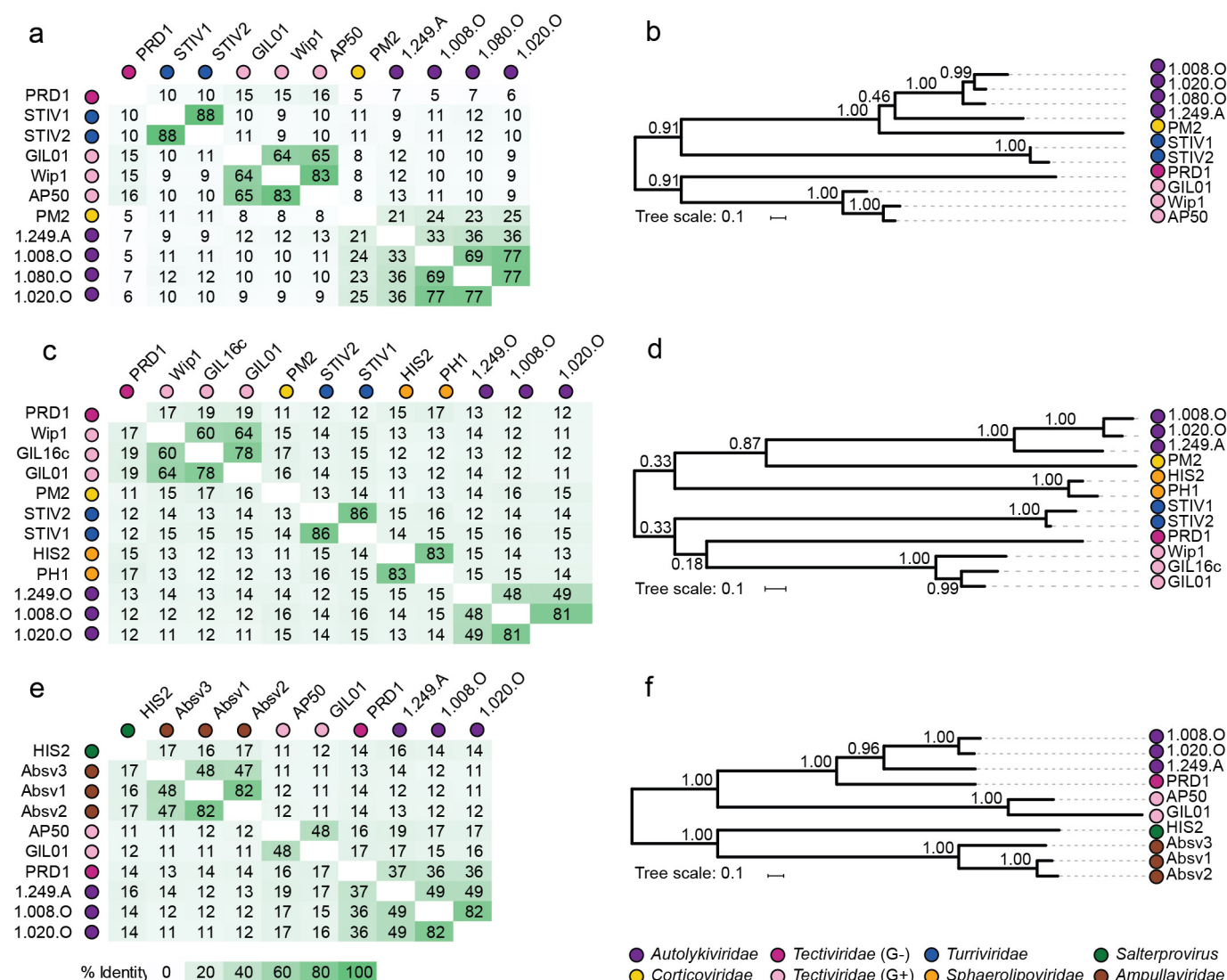
Extended Data Figure 2 | Whole-genome alignments show that the family Autolykiviridae consists of five major sequence diversity clusters. **a**, Maximum Likelihood phylogeny of whole-genome nucleotide alignments of 21 autolykiviruses. Alignments were made with Clustal Omega and the phylogenetic tree was generated with PhyML-SMS with aLRT branch supports. Scale bar, substitutions per base. **b**, Percentage of whole-genome nucleotide identities among 21 autolykivirus genomes on

the basis of the Clustal Omega alignment. Assumptions of 50% and 95% identity for genus and species classifications⁸³, respectively, suggest that these viruses represent two genera (groups A, B, C, D and group E) and five species. Two viruses with identical genomes were isolated at time points 39 days apart (1.048.O and 1.102.O), viruses with the same number and different letter suffixes represent lineages derived from a single plaque that gave rise to variable morphotypes during serial purification.



Extended Data Figure 3 | Genomes of members of the *Autolykiviridae* are syntenic despite extensive diversity at the nucleotide level. Virus genomes are grouped by nucleotide similarity (as identified in Extended Data Fig. 2). Homologous proteins were identified by performing an all-by-all BLASTp, requiring a minimum bitscore of 50, and clustering all pairs unweighted, using MCL with an inflation parameter set to 1.4 (Methods), cluster membership is identified by the label over the block arrows in the genome diagram. Protein clustering reveals that in

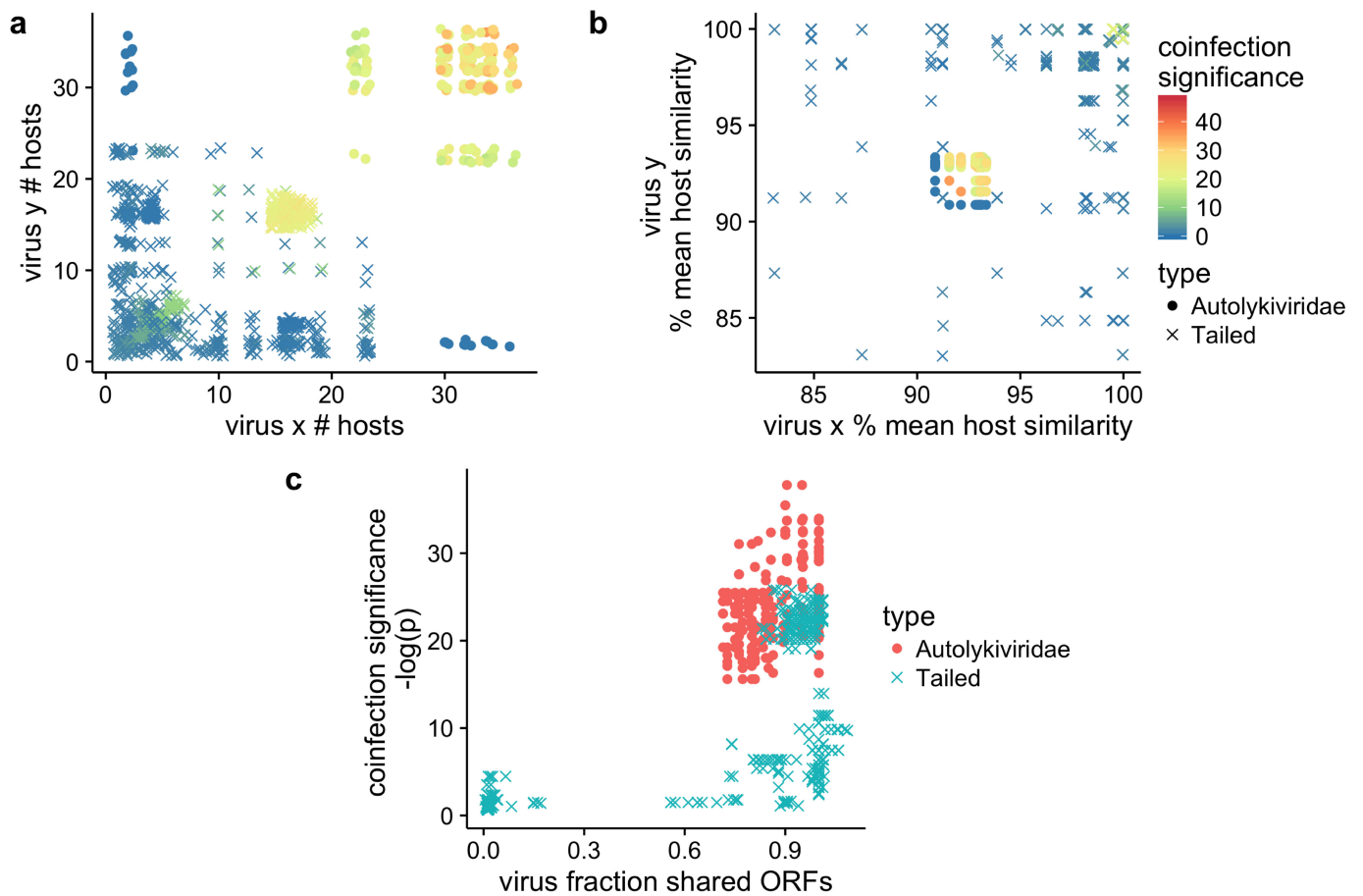
addition to the six proteins identifiable by sequence similarity as core to all characterized autolykiviruses, additional protein clusters are shared among various subsets of the identified viral genome groups. For example, in the region of the genome to the right of the major capsid protein, 17 out of 18 viruses (genome groups A, B, C and D) share a set of seven protein clusters of unknown function (c11, c12, c13, c14, c15, c16 and c17); among these viruses, two additional proteins are shared only within subsets of the genomes (c26 in genome groups A and B; c19 in genome groups C and D).



Extended Data Figure 4 | Packaging and replication protein-sequence phylogenies of autolykiviruses are incongruent with respect to other known families of non-tailed dsDNA viruses. Autolykiviruses are most similar to the corticovirus PM2 in their major capsid protein, poorly resolved in their packaging ATPase, and most similar to the tectiviruses in their protein-primed DNA polymerase. Pairwise identities and phylogenies of the protein sequences of the DJR major capsid protein

(a and b), packaging ATPase (c and d) and protein primed DNA polymerase (e and f). Members of the *Tectiviridae* infecting Gram-positive and Gram-negative hosts are shown separately as G+ and G-, respectively. All alignments were performed using the ETE3 Toolkit with workflow eggNOG41. All trees are maximum-likelihood trees with aLRT branch supports.

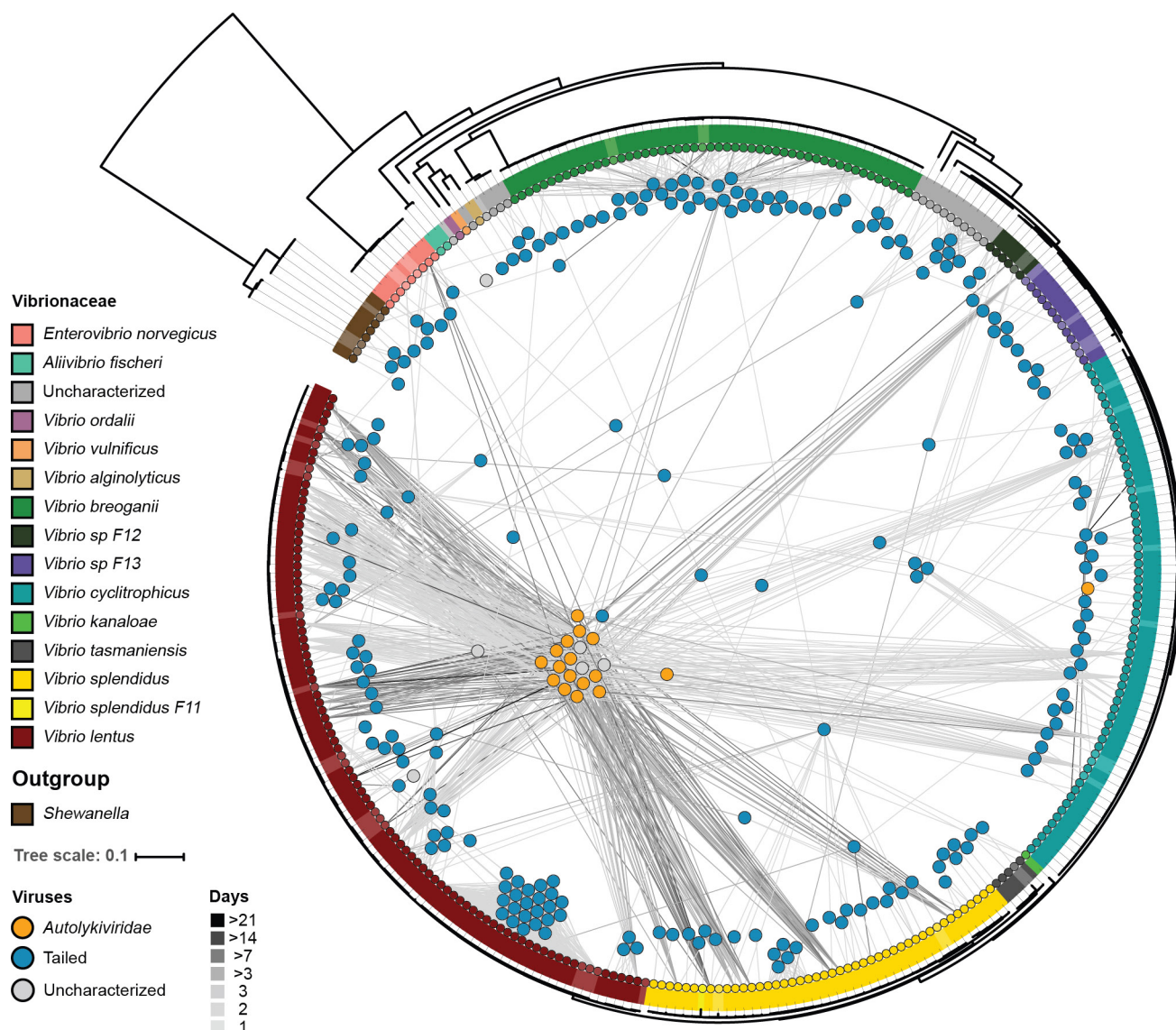
Autolykiviridae Tectiviridae (G-) Turriviridae Salterprovirus
Corticoviridae Tectiviridae (G+) Sphaerolipoviridae Ampullaviridae



Extended Data Figure 5 | Sequence-diverse autolykiviruses share extensively overlapping host ranges that include diverse hosts.

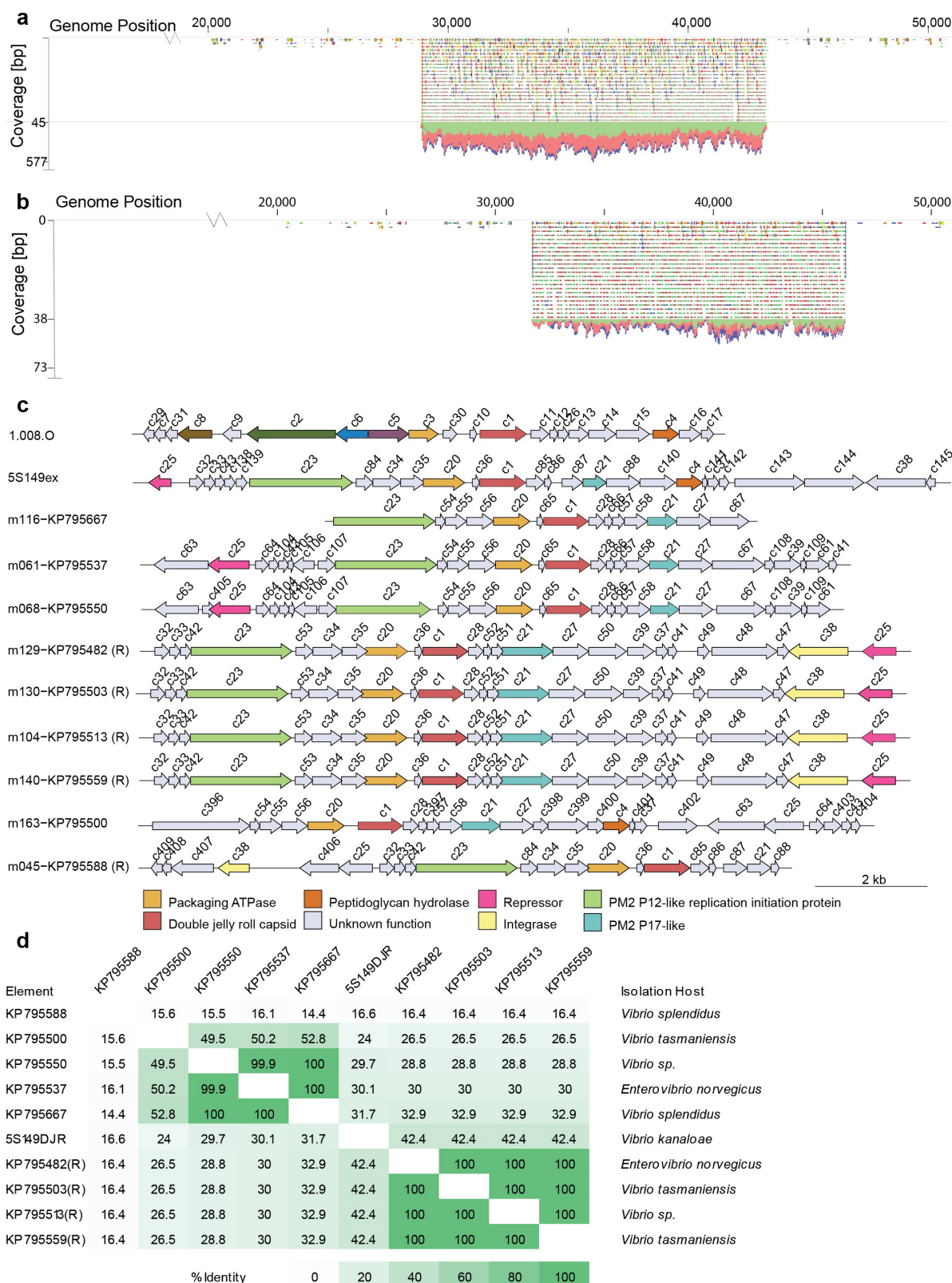
a, Pairwise coinfection significance by host count. Autolykiviruses exhibit highly significant host sharing. **b**, Pairwise coinfection significance compared to mean pairwise genomic similarity of the host. Autolykiviruses exhibit more significant host sharing than tailed phages of comparable host diversity. **a**, **b**, Coinfection significance as defined in

Methods. **c**, Pairwise coinfection significance compared to viral genomic similarity measured as a fraction of shared open reading frames (ORFs). Autolykiviruses exhibit more significant host sharing than tailed viruses of comparable genomic similarity. A total of 998 reciprocal pairs of tailed viruses and 236 reciprocal pairs of autolykiviruses are shown, representing all pairs of viruses within each group (141 unique tailed, 16 unique autolykiviruses) that share at least one host.



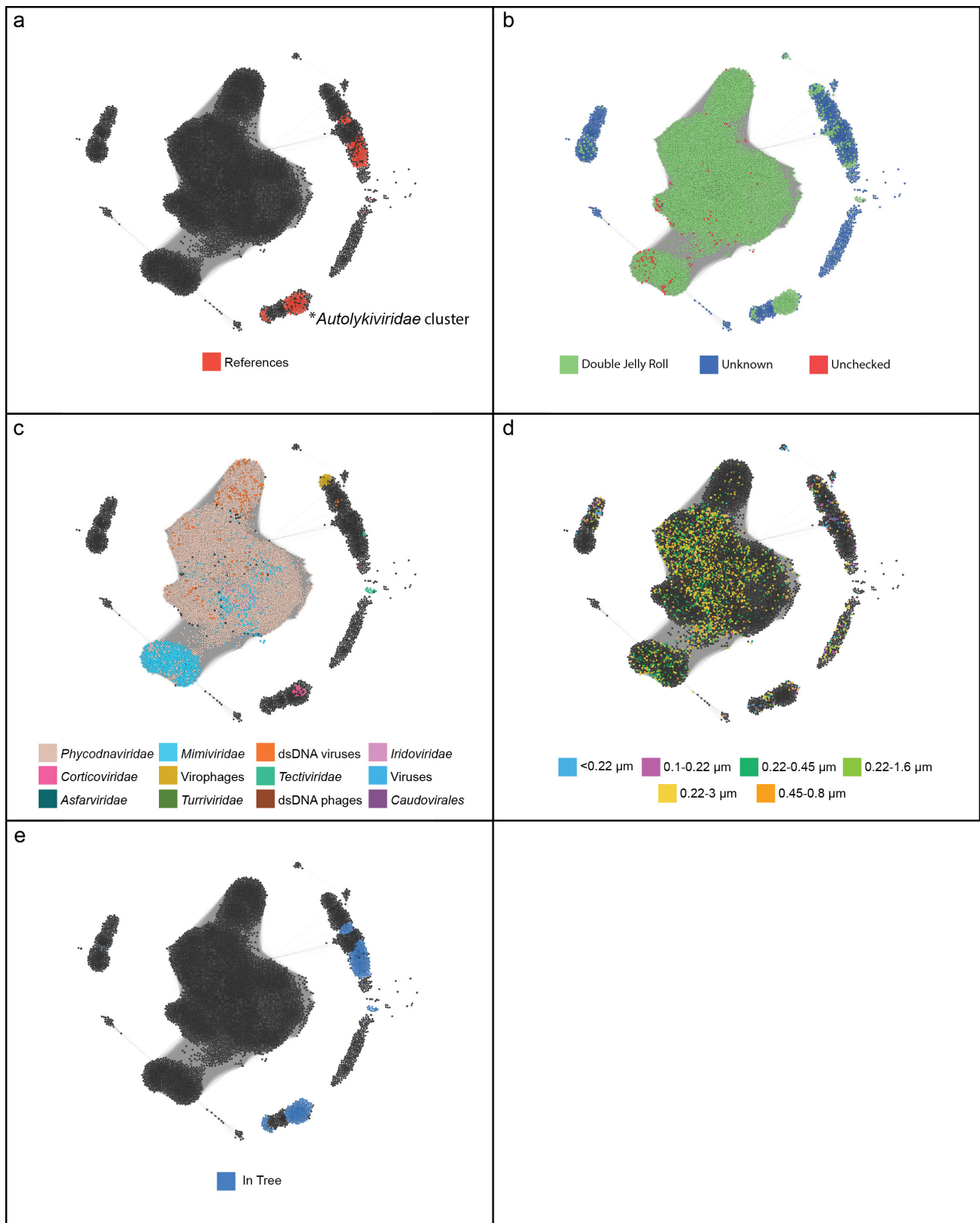
Extended Data Figure 6 | Autolykiviruses show delayed host lysis compared with other viruses. Inverted phylogenetic tree showing the relationships among all 318 assayed bacterial strains on the basis of the concatenated alignments of the *hsp60* and ribosomal protein genes, and using a partitioned model in RaxML to allow placement of 40 strains for which only the *hsp60* gene sequence was available (Methods). Isolates are generally non-clonal. Leaves represent Vibrionaceae isolates and are coloured by population (Methods). Nodes represent viruses and are

coloured by morphotype, as defined by major capsid protein or genome composition (Methods; non-tailed in orange, tailed in blue, unsequenced viruses in grey); edges represent infections with intensity increasing with increased time required for observation of plaques. Whereas 94% of tailed virus infections were detected within three days in host range assays, only 57% of autolykivirus infections were detected in that time, with 15% requiring more than seven days to be detected.



Extended Data Figure 7 | DJR elements in Vibrionaceae include naturally excising integrated prophages and broad host-range plasmids. Prophages of representative group 5 DJR elements (Fig. 4) naturally excise from their *Vibrio* hosts during growth in culture. Sequencing of nuclease-treated cell-free culture supernatants reveals sharply delineated regions of high coverage read mapping with respect to host genome background, indicating the presence of extracellular nuclease-protected prophage DNA. **a**, *V. kanaloae* 5S-149 DJR prophage. **b**, *Vibrio* 10N.286.55.C7 DJR prophage. **c**, Genome diagrams of the excising 5S-149 DJR prophage and the nine Vibrionaceae plasmids²⁸ that are identified here as DJR elements

show that they are syntenic and all share the DJR capsid protein, packaging ATPase and the corticovirus PM2 P17-like protein. MCL clustering of proteins on the basis of the BLASTp sequence similarity reveals that additional proteins, including integrases, repressors, peptidoglycan hydrolases and replication initiation genes, are common but not universal within these elements. **d**, Pairwise percentage of whole-genome nucleotide identities between 5S-149 DJR prophage and the DJR Vibrionaceae plasmids show that these elements are highly diverse at the nucleotide level and that 100% nucleotide-identical 13.6-kb plasmids are found in hosts in multiple species.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Network of DJR virus capsids identified in bacterial and archaeal genomes and marine metagenomes. Iterative HMM-based searches of marine metagenomes, on the basis of a reference panel of autolykiviruses and previously identified DJR capsid bacterial and archaeal viruses, yield approximately 15,000 proteins following stringent quality control filtering of the initial approximately 45,000 sequences that were recovered. Network visualization reflects MCL clustering of BLASTp-based similarities among sequences. **a**, Placement of reference panel sequences within the network. **b**, Characterization of proteins as DJRs on the basis of sequence- and structural-similarity-based annotation. **c**, Best BLASTp matches to RefSeq viruses, bitscore requirement of 50. **d**, Association of *Tara* Oceans-derived sequences to size fraction of isolation. **e**, Subset of sequences selected for phylogenetic analyses (Fig. 4) on the basis of membership in protein clusters strongly supported as bacterial and archaeal virus DJR capsids and requiring a length of ≥ 200 amino acids (Methods). We note that this selection is conservative, given

the greater number and diversity of sequences recovered by our HMM-based search that passed all quality controls and show no structural- or sequence-based similarity to any other proteins, and thus were excluded from further analyses. The observed dominance of eukaryotic virus DJR capsids in this search is predicted to reflect four major aspects of our approach. First, inclusion of cellular metagenomes allows capture of large viruses such as the *Mimiviridae* (>400 nm), *Iridoviridae* (120–350 nm) and *Phycodnaviridae* (100–220 nm). Second, some *Phycodnaviridae* have been shown to encode up to eight sequence-diverse copies of their DJR major capsid gene⁸⁴. Third, <0.22 μ m viral metagenomes are biased against recovery of bacterial and archaeal DJR viruses, as described here. And fourth, the sequence content of HMMs using iterative searches is defined by the search space, such that if eukaryotic virus DJR capsid sequences are well represented, as they are in the larger size-fraction sequence databases used here, they will drive searches towards increased detection of similar sequences.

Extended Data Table 1 | Metagenomes used in this study

Metagenomic dataset	Data Source
<i>Tara</i> Oceans Viromes ⁸⁵	ftp://ftp.imicrobe.us/projects/197/TOV_43_all_contigs_predicted_proteins.faa.gz
<i>Tara</i> Oceans OM-Reference Gene Catalog ⁵⁰	ftp://ftp.sra.ebi.ac.uk/vol1/ERA412/ERA412970/tab/OM-RGC_seq.release.tsv.gz
Methane Seep Sediment	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA290197
Rifle Sediment ^{86–88}	https://www.ncbi.nlm.nih.gov/bioproject/?term=288027
Mediterranean Sea Virome ⁸⁹	GenBank Accessions: AP013358:AP014505
Mediterranean Sea Metagenome ⁹⁰	GenBank Accessions: GU942957:GU943153
Chesapeake Bay Virome	https://www.ncbi.nlm.nih.gov/sra/SRR4293227/
NCBI environmental metagenomes ⁹¹	ftp://ftp.ncbi.nlm.nih.gov/blast/db/env_nr*.tar.gz
Nahant Light Fraction Viral Metagenome	This paper, GenBank Accession: PDMW00000000
Nahant Viral Metagenome	This paper, GenBank Accession: PDMX00000000

Description and data sources^{50,85–91} for each of the metagenomic datasets used in this study.

Extended Data Table 2 | Contigs of DJR elements

Group	Type	Host phylum	Accession Number	Element or Contig Name
1	<i>Autolykiviridae</i>	Proteobacteria (c: gamma)	In Supplementary Data File 3	1.008.O
2	<i>Corticoviridae</i>	Proteobacteria (c: gamma)	NC_000867.1	PM2
3	Host-associated	Zixibacteria	MEWP01000034.1 (R)	Zixibacteria bacterium RBG_16_53_22 RBG_16_scaffold_15471
4	Metagenomic Contig		LAZR01011096.1	LCGC14_contig011102
5	Host-associated	Proteobacteria (c: gamma)	AJYX02000003.1: 28913-43245	Excised <i>Vibrio kanaloae</i> 5S-149 DJR prophage
6	Metagenomic Contig		CEUD01119479.1	TARA_137_MES_0.22-3_scaffold209516_1
7	Host-associated	Proteobacteria (c: alpha)	JPUR01000104.1 (R)	Marinosulfonomonas sp. PRT-SC04 contig_12486
8	Metagenomic Contig		newLF_contig_7959 (R)	Nahant-LF_contig7959
9	Metagenomic Contig		LF_contig_41867	Nahant-LF_contig41867
10	Metagenomic Contig		LAZR01031772.1	LCGC14_contig031806
11	Metagenomic Contig		CEPX01414959.1	TARA_037_MES_0.1-0.22_C20701301_1
12	Metagenomic Contig		CEPX01063969.1	TARA_037_MES_0.1-0.22_scaffold154415_1
13	Metagenomic Contig		LAZR01001928.1	LCGC14_contig001928
14	Metagenomic Contig		LAZR01007575.1	LCGC14_contig007576
15	<i>Turriviridae</i>	Crenarchaeota	NC_005892.1	STIV1
16	<i>Tectiviridae</i> (G-)	Firmicutes	NC_011523.1	AP50
17	Host-associated	Actinobacteria	CP006261.1	<i>Streptomyces collinus</i> Tu 365 plasmid pSCO2
18	Host-associated	Acidobacterium	MEKI01000026.1	Acidobacteria bacterium RBG_13_68_16 RBG_13_scaffold_1666
20	<i>Tectiviridae</i> (G-)	Proteobacteria (c: gamma)	NC_001421.2	PRD1
21	Metagenomic Contig		LAZR01015278.1 (R)	LCGC14_contig015288
22	Metagenomic Contig		CEPX01030804.1 (R)	TARA_037_MES_0.1-0.22_scaffold77974_1
23	Metagenomic Contig		CBAY_603174 (R)	Chesapeake Bay Virome contig_603174
24	Metagenomic Contig		CBAY_52461	Chesapeake Bay Virome contig_52461
25	Host-associated	Thaumarchaeota	CP007174.1: 2808000-2828000	<i>Nitrososphaera evergladensis</i> SR1
26	Host-associated	Thaumarchaeota	CP002408.1: 258549-274228	<i>Nitrososphaera gargensis</i> Ga9.2
27	Host-associated	Crenarchaeota	CP003317.1	<i>Pyrobaculum oguniense</i> TE7 extrachromosomal element
28	Metagenomic Contig		lcl_contig_23053 (R)	Chesapeake Bay Virome contig_23053
29	Metagenomic Contig		CEPX01198169.1 (R)	TARA_037_MES_0.1-0.22_scaffold345758_1

Source information for contigs of DJR group representatives presented in Fig. 4b. Additional notations in 'Accession number' columns include: (1) coordinate information if a contig represents an extraction from a larger sequence; (2) an R if the contig is presented in the reverse orientation with respect to annotations provided in Supplementary Table 1.

Tet2 promotes pathogen infection-induced myelopoiesis through mRNA oxidation

Qicong Shen^{1*}, Qian Zhang^{1,2*}, Yang Shi³, Qingzhu Shi³, Yanyan Jiang¹, Yan Gu¹, Zhiqing Li³, Xia Li², Kai Zhao², Chunmei Wang², Nan Li¹ & Xuetao Cao^{1,2}

Varieties of RNA modification form the epitranscriptome for post-transcriptional regulation¹. 5-Methylcytosine (5-mC) is a sparse RNA modification in messenger RNA (mRNA) under physiological conditions². The function of RNA 5-hydroxymethylcytosine (5-hmC) oxidized by ten-eleven translocation (Tet) proteins in *Drosophila* has been revealed more recently^{3,4}. However, the turnover and function of 5-mC in mammalian mRNA have been largely unknown. Tet2 suppresses myeloid malignancies mostly in an enzymatic activity-dependent manner⁵, and is important in resolving inflammatory response in an enzymatic activity-independent way⁶. Myelopoiesis is a common host immune response in acute and chronic infections; however, its epigenetic mechanism needs to be identified. Here we demonstrate that Tet2 promotes infection-induced myelopoiesis in an mRNA oxidation-dependent manner through Adar1-mediated repression of Socs3 expression at the post-transcription level. Tet2 promotes both abdominal sepsis-induced emergency myelopoiesis and parasite-induced mast cell expansion through decreasing mRNA levels of Socs3, a key negative regulator of the JAK–STAT pathway that is critical for cytokine-induced myelopoiesis. Tet2 represses Socs3 expression through Adar1, which binds and destabilizes Socs3 mRNA in a RNA editing-independent manner. For the underlying mechanism of Tet2 regulation at the mRNA level, Tet2 mediates oxidation of 5-mC in mRNA. Tet2 deficiency leads to the transcriptome-wide appearance of methylated cytosines, including ones in the 3' untranslated region of Socs3, which influences double-stranded RNA formation for Adar1 binding, probably through cytosine methylation-specific readers, such as RNA helicases. Our study reveals a previously unknown regulatory role of Tet2 at the epitranscriptomic level, promoting myelopoiesis during infection in the mammalian system by decreasing 5-mCs in mRNAs. Moreover, the inhibitory function of cytosine methylation on double-stranded RNA formation and Adar1 binding in mRNA reveals its new physiological role in the mammalian system.

During infection, sensing pathogen and inflammatory cytokines skews haematopoiesis towards myeloid development; however, the epigenetic mechanism for this pathogen infection-induced myelopoiesis is unclear⁷. We investigated the role of Tet2, a myeloid tumour suppressor, in pathogen infection-induced myelopoiesis. First, we subjected Tet2-deficient (knockout) and the littermate control (wild-type) mice to caecal ligation and puncture (CLP), a model of abdominal polymicrobial sepsis with acute mobilization and expansion of myeloid cells. Compared with the control mice, Tet2-deficient mice were protected from sepsis with lower mortality rates (Fig. 1a, see graph Source Data for statistics) and had lower clinical scores (Fig. 1b). The bactericidal activity of Tet2-deficient mice was not significantly affected within 24 h (Fig. 1c). One day after CLP, only the control mice developed obvious neutrophilia and inflammatory monocytosis,

whereas the neutrophil and monocyte numbers barely varied in Tet2-deficient mice (Fig. 1d). Significantly fewer peritoneal neutrophils and monocytes were also observed in Tet2-deficient mice (Fig. 1e). The increased neutrophil and monocyte numbers in control mice were associated with higher serum levels of inflammatory cytokines, such as tumour necrosis factor, keratinocyte chemoattractant and macrophage inflammatory protein-1 α , compared with Tet2-deficient mice (Fig. 1f). In a mouse model infected with parasite *Schistosoma japonicum* (Extended Data Fig. 1a), the numbers of mast cells in *Kit*^{W-sh/W-sh} mice transplanted with Tet2-deficient bone marrow cells were lower than those in the control group in the jejunum and ileum with diffuse mucosal inflammation and granulomatous reaction (Extended Data Fig. 1b, c). These data show that Tet2 promoted both emergency myelopoiesis, fuelling a cytokine storm during abdominal sepsis, and a long process of myelopoiesis, promoting expansion of tissue mast cells derived from bone marrow progenitor cells during chronic parasite infection.

We subjected significant varied genes in RNA sequencing (RNA-seq) data (Supplementary Table 1) from Tet2-deficient bone-marrow-derived mast cells (BMMCs) and control cells to Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis; and the expression variations of the genes in JAK–STAT and PI3K–AKT pathways, which are critical for inflammatory cytokine-induced myelopoiesis^{8,9}, were labelled near scatter plots (Fig. 2a). Among these genes, *Socs3*, a key suppressor of JAK–STAT signalling, was significantly increased in the Tet2-deficient group. Upregulation of *Socs3* in both BMMCs and Lin[−]Kit⁺ haematopoietic stem and progenitor cells (HSPCs) stimulated by interleukin-3 (IL-3), a critical cytokine for both acute infection-induced myelopoiesis⁸ and parasite infection-induced mast cell expansion⁹, was observed at both mRNA and protein levels in Tet2-deficient groups (Fig. 2b–d; see Supplementary Fig. 1 for gel Source Data). We also found decreased expression of IL-3 signal-induced genes in Tet2-deficient BMMCs (Extended Data Fig. 2a–c) and the impaired phosphorylation of Akt and STAT5 in both IL-3-stimulated BMMCs and bone marrow cells from Tet2-deficient mice (Extended Data Fig. 2d, e), further validating the defective JAK–STAT signalling in Tet2-deficient myeloid cells. Silencing of *Socs3* in Tet2-deficient BMMCs increased IL-3 signalling (Extended Data Fig. 2f), indicating that Tet2 promoted infection-induced myelopoiesis by repressing *Socs3* expression for efficient cytokine signalling.

To reveal the molecular mechanism of Tet2-mediated *Socs3* suppression, we first detected the DNA methylation levels of CpGs in two predicted CpG islands near the *Socs3* promoter, and found that all of these CpGs were hypomethylated in both wild-type and Tet2-deficient BMMCs (Extended Data Fig. 3a, b); loss of Tet2 still increased the mRNA level of *Socs3* when *de novo* transcription was inhibited (Extended Data Fig. 3c), indicating the potential role of Tet2 in regulating *Socs3* at post-transcriptional level.

¹National Key Laboratory of Medical Immunology & Institute of Immunology, Second Military Medical University, Shanghai 200433, China. ²Department of Immunology & Center for Immunotherapy, Institute of Basic Medical Sciences, Peking Union Medical College, Chinese Academy of Medical Sciences, Beijing 100005, China. ³Institute of Immunology, Zhejiang University School of Medicine, Hangzhou 310058, China.

*These authors contributed equally to this work.

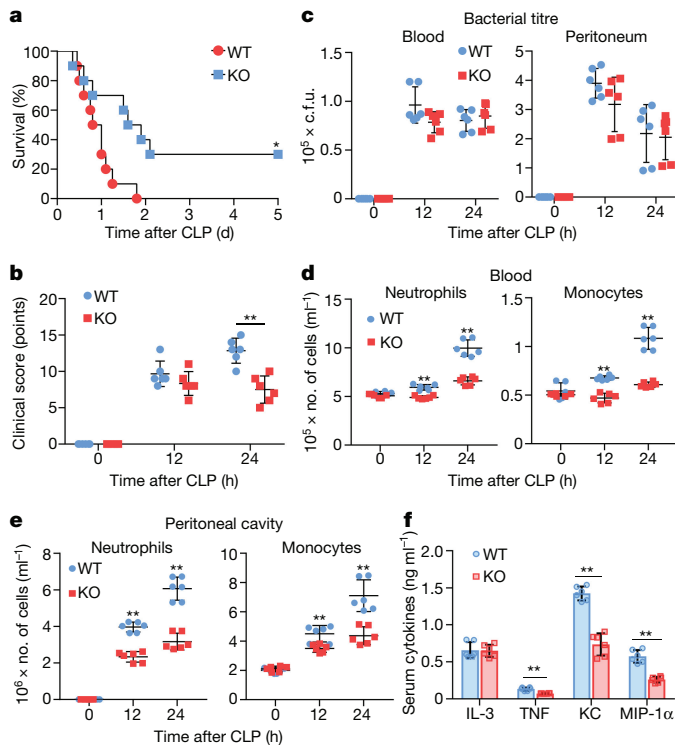


Figure 1 | Reduced emergency myeloopoiesis in Tet2-deficient mice in abdominal sepsis. a–f, Comparison of Tet2-deficient (KO) and the littermate control (WT) mice in the CLP model. a, Kaplan–Meier survival curve ($n = 10$). b, c, Clinical scores and bacterial titre ($n = 6$) (c.f.u., colony-forming units). d, e, Cell enumeration ($n = 6$). f, Levels of cytokines in serum 1 day after CLP ($n = 6$ biologically independent mice). TNF, tumour necrosis factor; KC, keratinocyte chemoattractant; MIP-1 α , macrophage inflammatory protein-1 α . * $P < 0.05$, ** $P < 0.01$ (Kaplan–Meier (a) or unpaired two-sided Student's t -test, mean and s.d. (b–f)). n , number of biologically independent animals.

Tet2 was recently identified as an RNA-binding protein, implying its general role in post-transcriptional regulation¹⁰. We performed three biological replicates of enhanced crosslinking and immunoprecipitation followed by high-throughput sequencing (eCLIP-seq)¹¹ for obtaining potential Tet2-binding RNAs in BMMCs (Extended Data Fig. 3d, e), and biological replicates correlated well with each other (Extended Data Fig. 3f). Among thousands of peaks identified in the three replicates, more than 80% were located in genic regions (Fig. 2e), including *Socs3* loci (Supplementary Table 2), and about 50–60% of CLIP peaks overlapped with each other in at least two replicates (Fig. 2f).

In RNA-seq data, we found more A-to-G mutations transcriptome-wide in the wild-type group than in the Tet2-deficient group, which contained an A-to-G mutation in the 3' UTR of *Socs3* (Supplementary Table 3). Most of the group-specific A-to-G mutants showed low mutation rates (Fig. 2g), and were more distributed in 3' UTRs in mature mRNA elements (Fig. 2h), consistent with RNA editing preferences in mRNAs. Furthermore, top motifs across the mutation sites identified by homer software¹² were found in the published motif across RNA A-to-I editing sites by adenosine deaminase Adar1 (ref. 13), implying a contribution of Adar1 to these mutations. Furthermore, the top motif in Tet2 CLIP peaks was also similar to the last seven bases of the published motif (Fig. 2i), implying a regulatory role of Tet2 in RNA editing.

Adar1 binds double-stranded RNA (dsRNA) and edits nearby adenosines¹³. Prediction of the secondary structure of the 3' UTR of *Socs3* showed that the A-to-G mutant was in a dsRNA stem loop (Extended Data Fig. 4a). The 3' UTR of *Lrrc47* bearing an A-to-G mutant in a predicted dsRNA stem in wild-type BMMCs was used as a classic control (Extended Data Fig. 4b). We validated that the

A-to-G mutation appeared in the 3' UTR of *Socs3* in wild-type BMMCs (Extended Data Fig. 4c). Adar1 binding in the 3' UTR of *Socs3* and *Lrrc47* in wild-type BMMCs was also validated (Fig. 3a and Extended Data Fig. 4d, e). Editing levels in the 3' UTR of *Socs3* gradually increased during differentiation of BMMCs (Extended Data Fig. 4f). Tet2 loss barely changed protein levels of Adar1, which located mainly in the nuclei of BMMCs (Extended Data Fig. 4g, h). Silencing of Adar1 indeed decreased editing rates of the sites in the 3' UTR of *Socs3* and *Lrrc47* (Fig. 3b and Extended Data Fig. 4i).

Silencing Adar1 increased mRNA and protein levels of *Socs3* in wild-type BMMCs (Fig. 3c, d). Equal signals were observed for the reporters bearing the A-to-G mutated and wild-type 3' UTR of *Socs3* (Fig. 3e). Furthermore, both wild-type Adar1 and an enzymatic activity mutant (Adar1^{E861A}), but not a dsRNA binding domain mutant (lacking N456–G743, Adar1 Δ dsRNA), repressed the signals of the reporter bearing the A-to-G mutated 3' UTR of *Socs3* (Fig. 3f). Previous studies indicated that Adar1 also mediates post-transcriptional regulation independent of RNA editing through other RNA-binding proteins^{14,15}. Through immunoprecipitation-coupled liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis as previously reported⁶, we found that Adar1 associated with two RNA-binding proteins (Fig. 3g), which were involved in regulating mRNA processing and stability^{16,17}. This implies that Adar1 post-transcriptionally represses *Socs3*, probably through its binding partner, which needs further investigation.

To investigate the regulatory role of Tet2 in Adar1-mediated inhibition of *Socs3* expression, we first found that silencing Adar1 in Tet2-deficient BMMCs barely increased the mRNA level of *Socs3* any further (Fig. 3h). Adar1 could inhibit both mRNA and protein levels of overexpressed *Socs3*, especially with the help of both wild-type Tet2 and a DNA-interacting residue mutant Tet2 (Tet2 Δ DNA)¹⁸, but not with an α -ketoglutarate (KG)-interacting mutant and the Fe²⁺ binding mutant of Tet2 (Tet2_{HxD}) (Fig. 3i and Extended Data Fig. 4j, k). Overexpression of Tet2 and Tet2 Δ DNA, but not Tet2_{HxD}, decreased the *Socs3* mRNAs in Tet2-deficient BMMCs (Fig. 3j). Although all three forms of Tet2 could bind *Socs3* mRNAs, a lower signal was observed in the Tet2_{HxD} mutant group (Fig. 3k), indicating that the stable catalytic structure also improves the interaction between Tet2 and *Socs3* mRNA. These results indicate that Tet2 promotes Adar1-mediated repression of *Socs3* mRNA in an enzymatic activity-dependent manner, but not in a DNA interaction-dependent manner.

Recently, 5-mC and its oxidation derivative, 5-hydroxymethylcytosine (5-hmC), catalysed by Tet proteins¹⁹, have been found in RNA in *Drosophila*³ and human cell lines⁴. We found that recombinant Tet2 could decrease 5-mC levels of transcribed RNAs *in vitro* in a substrate-dependent manner (Fig. 4a and Extended Data Fig. 5a, b). However, the major oxidation form of 5-mC *in vitro* in RNA was 5-hmC (Fig. 4b), unlike the oxidation forms in DNA (Extended Data Fig. 5c). Overexpression of wild-type Tet2 and Tet2 Δ DNA, but not Tet2_{HxD}, in HEK293T cells decreased 5-mC levels of mRNAs (Extended Data Fig. 5d–f). However, both overexpressing Tet2 in HEK293T cells and Tet2 knockout in BMMCs barely affected both the overall 5-mC levels of total RNAs and the levels of reported 5-mCs in tRNA^{Asp(GUC)} (ref. 20) (Extended Data Fig. 5g, h). Furthermore, silencing Tet2 in HEK293T cells and Tet2 knockout in BMMCs both increased 5-mC levels in mRNAs compared with the control cells (Fig. 4c and Extended Data Fig. 5i, j). However, both dot blot (800 ng mRNA) and LC–MS analysis (sensitivity 0.1 nM) barely detected signals of 5-hmC, 5-formylcytosine, (5-fC) and 5-carboxylcytosine (5-caC) in mRNAs in the tested cells (data not shown), indicating much lower levels of these oxidation products than 5-mC *in vivo*. These results indicate that Tet2 could decrease overall levels of 5-mCs in mRNAs in an oxidation-dependent manner.

To reveal 5-mC regulation by Tet2 transcriptome-wide, we performed bisulfite sequencing of mRNAs from Tet2-deficient BMMCs and the control cells. Both technical and biological replicates had good overlapping rates of identified 5-mCs, especially in Tet2-deficient group

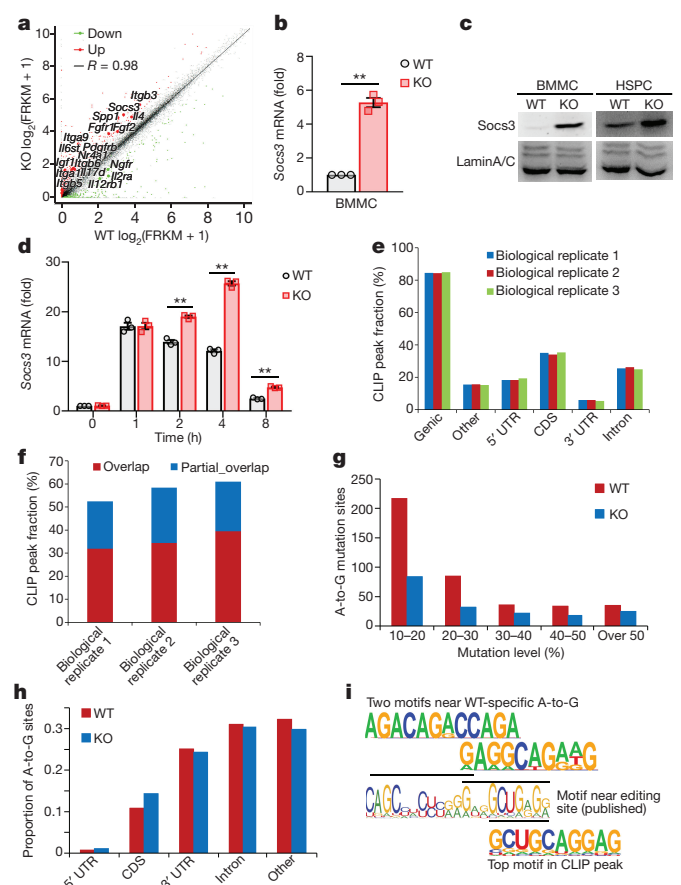


Figure 2 | Deficiency of Tet2 leads to increased transcripts and decreased A-to-I RNA editing for *Socs3*. **a**, Scatter plot of mRNA levels for a pair of Tet2-deficient (KO) and the control (WT) BMMCs. Upregulated (red) and downregulated (green) genes are coloured, varied genes in JAK-STAT and PI3K-AKT pathways are labelled near bigger plots. FRKM, fragments per kilobase of exon per million fragments mapped. **b–d**, qPCR and immunoblot assays of *Socs3* in BMMCs (**b**, **c**) and IL-3-stimulated Lin^{Kit} HSPCs for 8 h (**c**, **d**). **e**, **f**, Genomic distribution of CLIP peaks and their overlapping rates among the three biological replicates. **g**, **h**, Mutation levels and intragenic distribution of A-to-G mutants. CDS, coding sequence. **i**, Top motifs in 200 bp across A-to-G sites (top) and CLIP peaks (bottom), and the published motif (middle). * $P < 0.05$, ** $P < 0.01$, unpaired two-sided Student's *t*-test. Mean and s.e.m. of triplicate biological (**b**, **d**) replicates. Blots are representative of three independent experiments (**c**).

(Extended Data Fig. 6a, b). More methylcytosines were consistently identified, and with higher methylation levels in Tet2-deficient replicates than in controls (Fig. 4d and Extended Data Fig. 6c). There were many more group-specific methylcytosines in Tet2-deficient cells than in controls, more than half of which were located in the genic region, mostly in introns (Fig. 4e and Supplementary Table 4). As cell-specific alternative splicing leads to cell-specific intron retention, Tet2 may regulate 5-mC turnover in a cell-specific manner. Expression of genes bearing specific methylcytosines in the Tet2-deficient group or Tet2 CLIP peaks was more upregulated in Tet2-deficient cells (Extended Data Fig. 6d), implying an important role for Tet2-mediated mRNA oxidation in decreasing mRNA levels of genes such as *Socs3*. In mature mRNA-related elements, more methylcytosines located in the 3' UTR in the Tet2-deficient group than in controls (Fig. 4f), further confirming the regulatory role of Tet2 in the 3' UTR of mRNAs. When connecting the CLIP-seq data to specific methylcytosines in the Tet2-deficient group, we found that over 60% of genes or 3' UTRs bearing these methylcytosines were associated with CLIP peaks (Fig. 4g). Methylcytosines and CLIP peaks in exons located close to

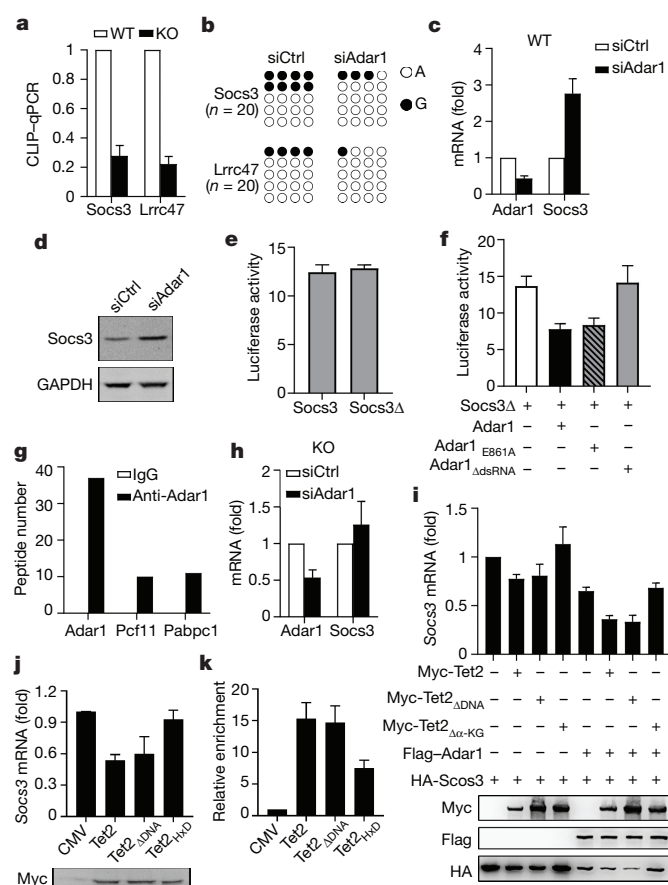


Figure 3 | Adar1 binds unmethylated *Socs3* mRNA and inhibits *Socs3* expression in a Tet2-dependent manner. **a**, qPCR analysis of editing regions in CLIPed RNAs for Adar1. **b–d**, **h**, RT-PCR-sequencing, qPCR and immunoblot assay of editing rates or gene expression in Adar1-silenced Tet2-knockout or wild-type BMMCs and the controls. **e**, **f**, Luciferase reporter assays of lysates of HEK293T cells transfected with pMIR plasmid with wild-type (*Socs3*) or mutant (*Socs3*Δ), and indicated plasmids for 24 h. **g**, Peptide numbers of potential Adar1 partners. **i**, qPCR and immunoblot analysis of HEK293T cells transfected for 24 h with indicated plasmids. **j**, **k**, Knockout BMMCs were transiently transfected with plasmids expressing the indicated forms of Tet2. *Socs3* mRNA levels were examined, and *Socs3* 3' UTR was from Tet2-immunoprecipitated RNA. Mean and s.d. of triplicate technical replicates (**a**, **c**, **e**, **f**, **h–k**). Blots are representative of three independent experiments (**d**, **i**, **j**).

each other in a mature mRNA view (Extended Data Fig. 6e). Among the 3' UTRs bearing several methylcytosines, *Socs3* was indeed in the list, and we validated that these methylcytosines appeared in Tet2-deficient cells (Extended Data Fig. 6f). We also found two other genes: *Zfp65* with an A-to-G mutant near methylcytosines in the 3' UTR where it was enriched in repeat elements; and *Tmed10* with a predicted stable dsRNA structure in the 3' UTR (Extended Data Fig. 6g–i). Adar1 also bound *Tmed10*, partly depending on Tet2 (Fig. 4h). We further validated increased RNA methylation levels in 3' UTRs of the three genes in Tet2-deficient BMMCs (Extended Data Fig. 6j).

We further investigated the underlying mechanism for the repressive role of 5-mC in Adar1 function, and found that overexpressed wild-type Tet2 and the Tet2_{ΔDNA} indeed decreased 5-mC levels of overexpressed *Socs3* mRNAs (Extended Data Fig. 7a). Oxidation forms of 5-mC in *Socs3* 3' UTR were not detected in wild-type BMMCs (Extended Data Fig. 7b). Mutating the 5-mCs in the 3' UTR of *Socs3* mRNA decreased the mRNA levels of *Socs3*, and overexpression of Tet2 barely synergistically repressed the mRNA levels of mutated *Socs3* with Adar1, compared with the wild-type *Socs3* (Extended Data Fig. 7c).

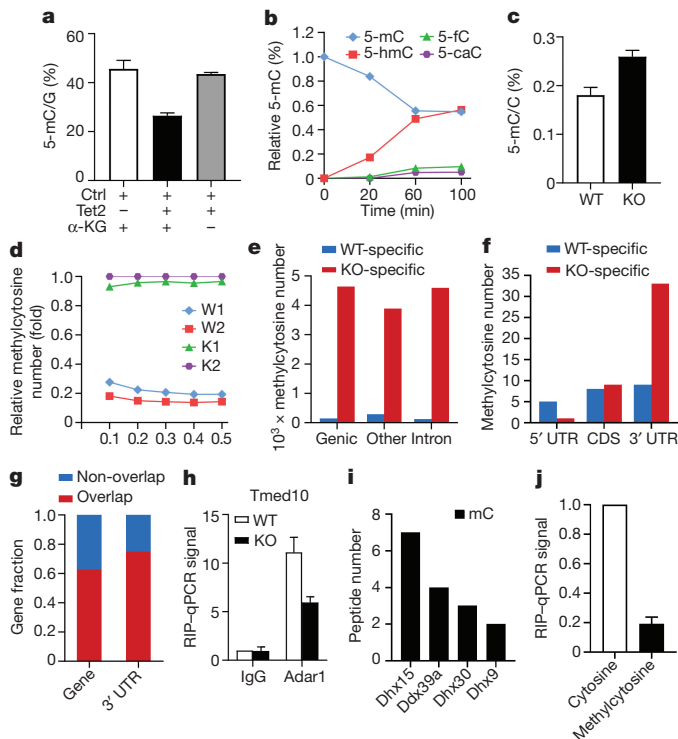


Figure 4 | Tet2 decreases mRNA methylation level in an oxidation-dependent manner for Adar1 targeting. a–c, LC–MS quantifying 5-mC and its oxides in Tet2-oxidized methylated RNAs *in vitro* (a, b) or in mRNAs (c). d, Relative methylcytosine numbers in replicate samples. e, f, Group-specific methylcytosine numbers in indicated genomic elements. g, Fractions of genes or 3' UTRs bearing specific methylcytosines in knockout group associated with CLIP peaks. h, Adar1-associated endogenous *Tmed10* 3' UTR. i, Peptide numbers of methylated *Socs3* 3' UTR-binding proteins in BMMCs. j, Adar1-associated methylcytosine- or cytosine-containing *Socs3* 3' UTR *in vitro*, normalized by the 1% input after incubation. Tet2-deficient BMMCs (KO) and the control (WT) were used (c–h). Mean and s.d. of triplicate technical replicates (a, c, h, j). RIP, RNA immunoprecipitation assay.

Overexpressed Adar1 preferentially bound the mutated form of the 3' UTR (Extended Data Fig. 7d). Furthermore, 5-mCs in the 3' UTR of *Socs3* mRNA could indeed repress the editing efficiency *in vitro* (Extended Data Fig. 7e). As dsRNA structure was essential for Adar1 binding, we found that loss of Tet2 decreased the dsRNA structure in *Socs3* mRNA (Extended Data Fig. 7f). DNA and RNA modifications can recruit specific readers to mediate epigenetic regulation. With the methylated 3' UTR of *Socs3* in pull-down assay, we identified several ATP-dependent RNA helicases (Fig. 4i), which are involved in altering RNA secondary structure and unwinding of dsRNA²¹. Adar1 indeed bound fewer methylated 3' UTRs of *Socs3* than unmethylated ones (Fig. 4j). These results imply that 5-mC in mRNA inhibits Adar1 function, probably through inhibiting dsRNA formation, which is essential for Adar1 binding.

As a type of RNA modification, 5-mC has been revealed to be important in regulating stability, protein translation and processing for tRNA, rRNA and even non-coding RNA and mRNAs^{22–27}. Our current study first revealed that Tet2 decreased 5-mC levels in mRNAs in an oxidation-dependent manner, and revealed a new function of increased 5-mC in mRNA owing to Tet2 loss in inhibiting the function of Adar1, especially in *Socs3* mRNAs (Extended Data Fig. 8a, b). Furthermore, our study revealed that Tet2 loss impaired dsRNA structure, which is essential for Adar1 function. With the data on transcriptome-wide editing sites that appeared in the wild-type control group but not the Tet2-deficient group, our study provided a general relationship between Tet2 and Adar1 function.

According to our data, amounts of 5-hmC were much lower than 5-mC in mRNA from the tested mammalian cells. Thus, additional enzymatic steps by an unknown protein may convert 5-hmC back to cytosine in mRNA. Further study will be needed to reveal this unknown mechanism, probably involving the members of the AlkB family²⁵. Moreover, our study implies that Tet2-mediated mRNA oxidation may be the critical step for RNA demethylation.

As a tumour suppressor gene, mutations of *Tet2* were largely found in myeloid malignancies and some solid cancers. For haematopoiesis, loss of Tet2 leads to increased abnormal myeloid cells in ageing⁷. And our study linked Tet2 to pathogen infection-induced myelopoiesis in an immunological way, and found that Tet2 strengthens cytokine signalling in myeloid differentiation by suppressing the repressor. Furthermore, mutations with a deficiency in enzymatic activity and downregulation of Tet2 in types of disease can lead to dysregulation of 5-mC in mRNA, which may be critically involved in the pathogenesis of myeloid disorders, such as myeloid tumours. This needs further investigation.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 22 November 2016; accepted 5 December 2017.

Published online 24 January 2018.

- Fu, Y., Dominissini, D., Rechavi, G. & He, C. Gene expression regulation mediated through reversible m⁶A RNA methylation. *Nat. Rev. Genet.* **15**, 293–306 (2014).
- Zhao, B. S., Roundtree, I. A. & He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat. Rev. Mol. Cell Biol.* **18**, 31–42 (2017).
- Delatte, B. *et al.* Transcriptome-wide distribution and function of RNA hydroxymethylcytosine. *Science* **351**, 282–285 (2016).
- Fu, L. *et al.* Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *J. Am. Chem. Soc.* **136**, 11582–11585 (2014).
- Álvarez-Errico, D., Vento-Tormo, R., Sieweke, M. & Ballestar, E. Epigenetic control of myeloid cell differentiation, identity and function. *Nat. Rev. Immunol.* **15**, 7–17 (2015).
- Zhang, Q. *et al.* Tet2 is required to resolve inflammation by recruiting Hdac2 to specifically repress IL-6. *Nature* **525**, 389–393 (2015).
- Moran-Crusio, K. *et al.* Tet2 loss leads to increased hematopoietic stem cell self-renewal and myeloid transformation. *Cancer Cell* **20**, 11–24 (2011).
- Weber, G. F. *et al.* Interleukin-3 amplifies acute inflammation and is a potential therapeutic target in sepsis. *Science* **347**, 1260–1265 (2015).
- Lantz, C. S. *et al.* Role for interleukin-3 in mast-cell and basophil development and in immunity to parasites. *Nature* **392**, 90–93 (1998).
- He, C. *et al.* High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells. *Mol. Cell* **64**, 416–430 (2016).
- Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
- Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Nishikura, K. A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.* **17**, 83–96 (2016).
- Anantharaman, A. *et al.* ADAR2 regulates RNA stability by modifying access of decay-promoting RNA-binding proteins. *Nucleic Acids Res.* **45**, 4189–4201 (2017).
- Sakurai, M. *et al.* ADAR1 controls apoptosis of stressed cells by inhibiting Staufen1-mediated mRNA decay. *Nat. Struct. Mol. Biol.* **24**, 534–543 (2017).
- Grzechnik, P., Gdula, M. R. & Proudfoot, N. J. Pcf11 orchestrates transcription termination pathways in yeast. *Genes Dev.* **29**, 849–861 (2015).
- Behm-Ansmant, I., Gatfield, D., Rehwinkel, J., Hilgers, V. & Izaurralde, E. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO J.* **26**, 1591–1601 (2007).
- Hu, L. *et al.* Crystal structure of TET2–DNA complex: insight into TET-mediated 5mC oxidation. *Cell* **155**, 1545–1555 (2013).
- Kohli, R. M. & Zhang, Y. TET enzymes, TDG and the dynamics of DNA demethylation. *Nature* **502**, 472–479 (2013).
- Goll, M. G. *et al.* Methylation of tRNA^{Asp} by the DNA methyltransferase homolog Dnmt2. *Science* **311**, 395–398 (2006).
- Jarmoskaite, I. & Russell, R. RNA helicase proteins as chaperones and remodelers. *Annu. Rev. Biochem.* **83**, 697–725 (2014).
- Popis, M. C., Blanco, S. & Frye, M. Posttranscriptional methylation of transfer and ribosomal RNA in stress response pathways, cell differentiation, and cancer. *Curr. Opin. Oncol.* **28**, 65–71 (2016).
- Amort, T. *et al.* Long non-coding RNAs as targets for cytosine methylation. *RNA Biol.* **10**, 1003–1008 (2013).

24. Hussain, S. *et al.* NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell Reports* **4**, 255–261 (2013).
25. Aas, P. A. *et al.* Human and bacterial oxidative demethylases repair alkylation damage in both RNA and DNA. *Nature* **421**, 859–863 (2003).
26. Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O. & Sorek, R. Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs. *PLoS Genet.* **9**, e1003602 (2013).
27. Luo, Y., Feng, J., Xu, Q., Wang, W. & Wang, X. NSun2 deficiency protects endothelium from inflammation via mRNA methylation of ICAM-1. *Circ. Res.* **118**, 944–956 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. L. Levine for providing Tet2 knockout mice, and K. Wang and C. Yi for helping with LC–MS analysis of RNA methylation. We thank C. Hu and W. Huang for technician support. This work was supported by grants from the National Natural Science Foundation of China (81788101, 31390431, 91542204, 31670884), the Shanghai Rising-Star Program

(17QA1405300) and the CAMS Innovation Fund for Medical Science (2016-12M-1-003).

Author Contributions X.C. designed and supervised the study. X.C., Q.Z. and Q.She. analysed the data and wrote the manuscript. Q.She. established pathogen infection mouse models. Q.She. and Q.Z. confirmed and genotyped mice, performed RNA methylation- and RNA editing-related experiments and analysed all the data of this study. Q.She. and Q.Z. performed CLIP, bisulfite sequencing and analysed the sequencing data. Y.S. performed the dot plot assays. Q.She. and Q. Shi constructed plasmids with aid from Q.Z. Y.J. performed parasite infections of mice. Y.G. and Z.L. sorted and analysed immune cells. X.L., K.Z., C.W. and N.L. provided reagents and advice.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to X.C. (caoxt@immunol.org).

METHODS

Mice and reagents. C57BL/6 mice were obtained from Joint Ventures Sipper BK Experimental Animal Company (Shanghai). Tet2-deficient mice on a C57BL/6 × 129/SvEv background were provided by R. L. Levine, and backcrossed to the C57BL/6 background in our laboratory. *Kit^{W-sh/W-sh}* mice were obtained from The Jackson Laboratory. All animal experiments were undertaken in accordance with the National Institute of Health Guide for the Care and Use of Laboratory Animals, with the approval of the Scientific Investigation Board of Second Military Medical University, Shanghai. Recombinant SCF and IL-3 were from PeproTech. Antibodies used were as follows: Anti-5-mC antibody (A3001, 10G4, lot: 2RC180672, Zymo Research), Anti-5-hmC antibody (Mab-31HMC, lot: 001, Diagenode), Anti-5-fC antibody (61227, lot: 34711001, Active Motif), Anti-5-caC antibody (61225, lot: 34711001, Active Motif), Anti-Adar1 antibody (A303-884A, Bethyl Labs), Anti-Socs3 (2923, lot: 2, Cell Signaling Technology), Anti-STAT5 (9363P, lot: 3, Cell Signaling Technology), Anti-STAT5 phosphorylated (Tyr694) (9359P, C11C5, lot: 4, Cell Signaling Technology), Anti-Akt (pan) (4685, 11E7, lot: 3, Cell Signaling Technology), Anti-Akt phosphorylated (Ser473) (4060, D9E, lot: 19, Cell Signaling Technology), Anti-ERK phosphorylated (Thr202/Tyr204) (4370, D13.14.4E, lot: 6, Cell Signaling Technology), Anti-ERK (9102, lot: 23, Cell Signaling Technology), Anti-JNK phosphorylated (Thr183/Tyr185) (4668, 81E11, lot: 9, Cell Signaling Technology), Anti-JNK (9258, 56G8, lot: 11, Cell Signaling Technology), Anti-Myc (2276, 9B11, lot: 24, Cell Signaling Technology), Anti-Flag (F1804, lot: SLBJ4607V, Sigma-Aldrich), Anti-HA (3724, lot: 3, Cell Signaling Technology), Anti-LaminA/C (4777, 4C11, lot: 9, Cell Signaling Technology), Anti-GAPDH (2118, 14C10, lot: 8, Cell Signaling Technology), Anti-Tet2 (MABE462, lot: Q2141878, Millipore), Anti-J2 monoclonal antibody (10010200, IgG2a, lot: No.J21611 Scions), Anti-β-actin (3700, 8H10D10, lot: 13, Cell Signaling Technology), Anti-rabbit IgG-HRP (31460, lot: RB230194, Thermo Fisher Scientific), Anti-mouse IgG-HRP (31430, lot: RA230188, Thermo Fisher Scientific), Anti-biotin (ab6643, lot: GR99377-4, Abcam), APC anti-Kit (17-1171-81, 2B8, lot: E07202-1631, eBioscience), PerCP cyanine5.5 anti-CD11b (45-0112-82, M1/70, lot: 4301974, eBioscience), PE cyanine7 anti-FcεRI (25-5898-80, MAR-1, lot: E16926-105, eBioscience), FITC anti-Ly6G (127606, 1A8, lot: B164314, Biolegend), Alexa Fluor 700 anti-Ly6C (128024, HK1.4, lot: B224165, Biolegend), BV421 anti-F4/80 (123137, BM8, lot: B202010, Biolegend). All the antibodies are all commercially available and validated by the suppliers according to the validation statements on the manufacturers' websites.

Cell purification and cultures. For BMMCs, mouse bone marrow cells were isolated from femurs and cultured in RPMI1640 medium plus 10% (v/v) FBS (Gibco) with 10 ng ml⁻¹ IL-3 and 5 ng ml⁻¹ SCF. Half of the medium was replaced by fresh medium with cytokines every 2–3 days during culture. After 4–6 weeks in culture, BMMCs were stained to confirm the surface expression of FcεRI and Kit. Cells with purity greater than 97.5% were used for subsequent experiments. For HSPCs, bone marrow cells were stained with anti-Lineage-PE and anti-Kit-APC antibodies, and the HSPCs (purity > 99%) were sorted using a MoFloXDP High-performance Cell Sorter (DACO Cytomatix). Cells were defined as monocytes (Ly6G⁻Ly6C^{high}CD11b^{low}), neutrophils (Ly6G⁺Ly6C^{int}CD11b⁺F4/80⁻), HSPC (Kit⁺Lin⁻) and mast cells (FcεRI⁺Kit⁺). HEK293T cells were from American Type Culture Collection and cultured in endotoxin-free DMEM (Thermo Fisher Scientific) supplemented with 10% (v/v) FBS (Thermo Fisher Scientific) without further authentication and mycoplasma contamination test.

Animal models. CLP. Six-week-old mice were used in this study. The rodent model of sepsis was performed as previously described²⁸. All experiments included age-matched controls. To induce mid-grade CLP, approximately 30–50% of the caecum was ligated. To induce high-grade CLP, approximately 60–80% of the caecum was ligated. Only experiments testing survival used high-grade CLP. The clinical score of animals was assessed as previously described⁸.

Parasite infection. Female 6-week-old mast-cell-deficient (*Kit^{W-sh/W-sh}*) mice had been reconstituted with bone marrow cells from the donor mice. Then, each of the mice was infected with *S. japonicum* by skin contact with 20 cercariae, with at least 6 mice in each group. Successful infection was confirmed by the detection of parasite eggs by stool examination before use in the subsequent experiments.

RNA quantification. RNA was extracted with TRIzol reagent (Thermo Fisher Scientific) and reversed-transcribed with a reverse transcription system (Toyobo). Reverse transcription products of different samples were amplified by a LightCycler System (Roche) using the SYBR Green PCR Premix Ex Taq (Takara) according to the manufacturer's instructions, and data were normalized by the level of β-actin or 1% input RNAs in each individual sample. The 2^{-ΔΔC_t} method was used to calculate relative expression changes. With the help of dissociation curve analysis and the sequencing of PCR products, pairs of specific primers of each cDNA were designed and selected, without any primer dimers or unspecific amplification detected. The sequences of the primers for quantitative RT-PCR are in Extended Data Table 1.

Biochemical assay of Tet2-mediated oxidation of 5-mC in RNA. The Tet2-mediated RNA oxidation assay was conducted with the experimental workflows as previously reported²⁹, with the presence of RNasin Plus RNase Inhibitor (Promega). A reaction mixture contained 250 ng of 5-mC-bearing 3' UTR of *Socs3*, the catalytic domain of recombinant human Tet2 (active motif) or immunoprecipitated Tet2 mutants overexpressed in HEK293T cells, oxidation reagent 1 (1.5 mM Fe(NH₄)₂(SO₄)₂·6H₂O), with (α-KG+ group) or without (α-KG- group) oxidation reagent 2 (333 mM NaCl, 167 mM HEPES (pH 8.0), 4 mM ATP, 8.3 mM DTT, 3.3 mM α-KG and 6.7 mM L-ascorbic acid). The 3' UTRs of *Socs3* were obtained from *in vitro* transcription by using T7 RNA polymerase (Thermo Fisher Scientific). The reaction was incubated at 37 °C for 60 min in a thermocycler and the enzyme was removed immediately afterwards by an RNeasy MinElute Cleanup Kit (Qiagen) according to the manufacturer's instructions. DNase-treated and purified RNAs were used for subsequent analysis.

Quantification of 5-mC by LC-MS/MS. Quantification of 5-mC by LC-MS/MS was performed as previously reported³⁰, with the following modifications: 0.5 U rSAP (NEB) was used instead of alkaline phosphatase (Sigma-Aldrich). The mass transitions of *m/z* 258.0–126.1 (5-mC), *m/z* 274.0–142.1 (5-hmC), *m/z* 271.2–140.1 (5-fC), *m/z* 286.1–156.1 (5-caC) and *m/z* 244.1–112.0 (cytosine, C) were monitored and recorded, a series of concentrations of pure authentic nucleoside standards (C, from 50 nM to 2,000 nM, Sigma-Aldrich; 5-mC, from 0.5 nM to 50 nM, Sigma-Aldrich; 5-hmC, Santa Cruz, 5-fC and 5-caC, Berry & Associates, from 0.1 nM to 50 nM) were run for every batch of experiments to obtain their corresponding stand curves. Concentrations of nucleosides in mRNA samples were deduced by fitting the signal intensities into the stand curves. The ratios of 5-mC/G or 5-mC/U were subsequently calculated. Relative oxidation amounts were compared with 5-mC/U₀, 0 min as 100%.

Dot blot. Dot blot assays for 5-mC, 5-hmC and 5-caC quantification in 90 °C denatured RNAs or DNAs were conducted as previously reported³. In brief, RNAs or DNAs were spotted onto a nylon membrane (GE Healthcare). The membrane was dried and crosslinked twice with 200,000 μJ cm⁻² ultraviolet light. The membrane was blocked in 5% BSA in PBS + 0.1% Tween-20 for 1 h before transfer into blocking solution supplemented with 5-mC or other modification antibody and incubated overnight at 4 °C. After secondary antibody incubation and wash, dot blots were visualized using SuperSignal West Femto Chemiluminescent Substrate (Thermo Fisher Scientific) by a chemiluminescent imaging system. The same amounts of denatured DNAs were degradation by TURBO DNase (Thermo Fisher Scientific) as negative control. To remove possible contaminating genomic DNAs, all RNAs samples were treated by the DNase.

Plasmid constructs. Full-length mouse *Socs3* or *Adar1* was PCR amplified using reverse transcribed RNAs from BMMCs. Tet2 eukaryotic expression vector was obtained as previously reported⁶. Mouse Tet2 mutant forms were as follows: Tet2_{HSD}, HS(H/Y)R(D/A)QQ; Tet2_{Δα-KG}, T(R/M)I(S/F)LVLYRH, CTN(R/G)RCSQN; and Tet2_{ΔDNA}, (W/R)SMYYNGC(K/E)FAR(S/N). They were generated by PCR-based amplification of the construct coding the wild-type protein and subcloned into the pCMV-Myc-N (Clontech). Wild-type *Adar1*, *Adar1*_{E861A} and *Adar1*ΔdsRNA³¹ (delete N456–G743) amplicons were subcloned into the pcDNA3.1-Flag-C (Invitrogen) vector. The *Socs3* and *Socs3*_{C-to-G} (chromosome 11: 117967529, 522, 518, 507, 485) full length were constructed into pcDNA3.1-HA-N (Thermo Fisher Scientific) vector. The 3' UTR of *Socs3* and the editing mutants *Socs3*Δ (chromosome 11: 117967036) was subcloned into pMIR-REPORT Luciferase (Thermo Fisher Scientific) vector. All constructs were confirmed by DNA sequencing.

eCLIP-seq. Biological replicates of BMMCs that had different culture start dates and crosslinked end dates were collected. eCLIP was conducted as previously reported¹¹, with the following modifications. Ultraviolet-crosslinked (first 400 mJ cm⁻² and then again at 200 mJ cm⁻²) BMMCs (4 × 10⁷) were lysed in lysis buffer (50 mM Tris-HCl (pH 7.4), 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate and protease inhibitors) and sonicated. Lysates were treated with RNase I (Thermo Fisher Scientific) to fragment RNA and DNase I to remove DNA, while Tet2 antibody was bound to dynabeads in lysis buffer for 1 h at room temperature. Beads were washed three times using lysis buffer and incubated with proteins lysates for 4 h at 4 °C. A 3' RNA adaptor was ligated onto the RNA with T4 RNA ligase (NEB). Protein–RNA complexes were run on a 4–12% gradient Bis-Tris Gels (Invitrogen), transferred to PVDF membranes, and RNA was isolated off the membrane identically to standard eCLIP. A fraction of sample was used for western blot of CLIPed endogenous Tet2, indicating regions excised for eCLIP library preparation. After purification, RNAs were reverse transcribed with Superscript III reverse transcriptase (Invitrogen) with nested specific primer and a protected reverse PCR primer as previously reported³², free primer was removed (ExoSap-IT, Affymetrix) and a 3' DNA adaptor was ligated onto the cDNA product with T4 RNA ligase (NEB). Libraries were then amplified with Premix Taq PCR mix (Takara). Adapters and primers were designed according to the commercial indexing and sequencing

primers. Sequencing reads were processed and mapped according to eCLIP procedure. Peaks were identified using a 'valley seeking' algorithm, in which a peak is called if the valley of certain depth are found on both sides³³. Peaks were filtered with peak height above 5 as cut off. Enriched motifs in CLIP peaks were analysed by homer software using strand-specific sequences from peak regions as inputs.

RIP assay. RIP was conducted as previously reported³⁴. For endogenous RIP, cell lysates were made from wild-type or Tet2 knockout BMMCs in polysome lysis buffer (100 mM KCl, 5 mM MgCl₂, 10 mM HEPES pH 7.0, 0.5% NP40) supplemented with DTT, protease inhibitor cocktail (Roche) and RNase Inhibitor (Promega) on ice. Lysates were sonicated and stored at -80°C . The Protein G magnetic beads were pre-blocked for 1 h with rotation in NT2 buffer supplemented with 5% BSA (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM MgCl₂, 0.05% NP40). The antibody was then added for 4 h at 4°C with rotation, followed by wash antibody-coated beads with 1 ml of ice-cold NT2 buffer four or five times. The cell lysate was thawed on ice and the insolubles were removed by centrifugation at 4°C ; 1% of total lysate was saved for input. Incubated cleared lysate and antibody were mixed for 2 h at 4°C with rotation. Beads were washed four times with NT2 buffer and the RNA released by proteinase K for 30 min at 55°C . Then RNA was isolated by adding TRIzol to the beads and glycogen (Thermo Fisher Scientific) added as a carrier to aid the precipitation reaction. For overexpression RIP experiments, the Myc-tagged Tet2 and its mutants, or haemagglutinin-tagged Socs3 or the mutant and Flag-tagged Adar1 were transfected into HEK293T cells or Tet2 knockout BMMCs for 36–48 h. Cells were collected in ice-cold polysome lysis buffer supplemented with protease inhibitor cocktail and RNase Inhibitor. For quantification in the RIP assay, cDNA amplicon signals were normalized by those from the reserved 1% input and used for comparison between experimental samples. When antibody against target was initially used in one type of cell, normalized signals in IgG groups were used for comparison with experimental samples. RNAs from both the reserved 1% input and the immunoprecipitated samples were treated with TURBO DNase, purified by TRIzol LS Reagent, and then reverse transcribed using random hexamers and ProtoScript II Reverse Transcriptase (NEB). PCR was used to amplify the target regions. The qPCR primers for the assay of the association of Adar1 with Socs3 3' UTR are in Extended Data Table 1.

RNA-mediated interference. BMMCs and HEK293T cells were transfected with siRNA (20 nM) through use of INTERFERin reagent (Polyplus Transfection). The mouse-specific siRNAs targeting Socs3 and Adar1 were designed and synthesized by GenePharma (Shanghai). Sequences of siRNAs were as follows: Socs3 siRNA 5'-GCCUCAAUCACUUUUAUA-3'; Adar1 siRNA 5'-GCCUGCGAUAAAGCAUGAA-3'; Tet2 siRNA 5'-CCAUCA CAUUGCUUCUUU-3'.

J2 dsRNA pull-down. J2 antibody (1:50) was incubated with total nuclear extracts for 2 h at 4°C (Lysis buffer: 50 mM HEPES pH7.5, 50 mM KCl, 5 mM MgCl₂, 1 mM DTT, 1 \times complete protease inhibitors and RNase Inhibitor). Protein G-magnetic beads (Thermo Fisher Scientific) pre-blocked for 1 h with 1% BSA were added and incubated on a wheel for an additional 1.5 h at 4°C . dsRNA-antibody complexes were eluted and dsRNAs were extracted using TRIzol LS Reagent (Invitrogen). RNA treatment and qPCR analysis were performed as in the RIP assay. The qPCR primers for validation of the Socs3 mRNA are in Extended Data Table 1.

meRNA-IP. The unmethylated control RNA and modified control RNA were obtained from *in vitro* transcription by using T7 RNA polymerase (Thermo Fisher Scientific). The total RNA of wild-type and Tet2 knockout BMMCs was extracted using TRIzol reagent and added with the 1% controls. RNA immunoprecipitation using anti-5-mC or other modifications was performed as previously reported without fragmentation³⁵. In brief, 10 μg total RNA was incubated for 2 h at 4°C with 2 μg of affinity antibody in meRNA-IP buffer (150 mM NaCl, 0.1% NP-40, 10 mM Tris-HCl, pH 7.4). The mixture was then immunoprecipitated by incubation with Protein G beads (Thermo Fisher Scientific) at 4°C for an additional 1 h. After extensive washing using meRNA-IP buffer, bound RNA was eluted from the beads with TRIzol reagent and ethanol precipitated. RNA treatment and qPCR analysis were performed as in the RIP assay. The qPCR primers of the unmethylated and modified controls are in Extended Data Table 1. The PCR primers for Socs3 mRNA were the same as in the J2 dsRNA pull-down assay. Signals in the qPCR assay were calculated as in the RIP assay.

Bisulfite-PCR sequencing. DNAs and RNAs isolated from BMMCs were bisulfite-converted using an EZ DNA Methylation-Lightning Kit or an EZ RNA Methylation Kit, with modifications described in bisulfite sequencing (Zymo) according to the manufacturer's instructions. The RNAs were treated as in the RIP assay. The treated RNA was reverse transcribed using random hexamers and ProtoScript II Reverse Transcriptase (NEB). PCR was used to amplify the target regions. The PCR primers are in Extended Data Table 1. The amplicons were subcloned, and ten clones were selected and sequenced to calculate the methylation rate of each cytosine.

RNA-seq analysis. RNA-seq analysis was performed as previously reported⁶. TopHat and Cufflinks were used for data analysis with mm10 as reference³⁶. FRKM

(total fragments/(mapped reads (M) \times exon length (kilobases))) was calculated using cuffnorm for mRNA level quantification. HTSeq and DESeq were used for gene expression variation analysis (above twofold and $P < 0.05$, calculated by DESeq software on the basis of a negative binomial distribution)^{37,38}. Single-nucleotide mutation was analysed as previously reported^{39,40}. Filters were added for picking out mutations in mRNAs: (1) splice site information was added using a refGene.gtf file for excluding mutations there; (2) trimmed 10 nucleotides at both ends of sequencing reads, which could introduce biases; (3) with a quality score threshold of 25, a coverage threshold of 5 uniquely mapped reads; (4) a mutation supporting reads threshold of 2, and mutation rate between 0.1 and 0.95; (5) a mutation rate in DNA data under 0.05; (6) picking out group-specific mutants with mutation rates in one group three times higher than those in the other group.

RNA bisulfite high-throughput sequencing. Total RNAs were isolated by TRIzol reagent. Poly-A tailed RNAs were enriched by a GenElute mRNA Miniprep Kit (Sigma-Aldrich). The mRNAs were treated with TURBO DNase, purified by an RNeasy MinElute Cleanup Kit (Qiagen). Treated mRNAs with 0.5% unmethylated RNA controls from lambda genome were bisulfite-converted using an EZ RNA Methylation Kit, with some modifications (Zymo Research). In brief, 500 ng of mRNAs were converted using three cycles of 10-min denaturation at 70°C followed by 45 min at 54°C . RNA separation from bisulfite solution, desulfonation and purification were performed following the standard protocol of the kit. Strand-specific libraries were constructed following the 'TruSeq stranded mRNA sample preparation guide' from Illumina for 150PE sequencing using HiSeq3000. Clean data were processed as described in RNA-seq analysis. Read mapping and methylcytosine calling were analysed by BS-RNA⁴¹ using converted mm10 and lambda genomes as references. Splice site information was added using a refGene.gtf file during read mapping. Cytosines with a methylation rate above 0.1 and at least two reads supporting methylation in both technical and biological replicates of one group and below the non-conversion rate of each of the replicates of the other group (WT1 replicate 1: 0.0015; WT1 replicate 2: 0.0015; KO1 replicate 1: 0.0017; KO1 replicate 2: 0.0016; WT2 replicate 1: 0.0012; WT2 replicate 2: 0.0013; KO2 replicate 1: 0.0012; KO2 replicate 2: 0.0012) were chosen as group-specific cytosines.

***In vitro* RNA editing and Adar1 binding assays.** The 3' UTRs of Socs3 mRNAs were obtained from *in vitro* transcription by using T7 RNA polymerase (Thermo Fisher Scientific). Adar1-Flag-overexpressed HEK293T whole-cell lysates (1 mg) were incubated for 2 h at 4°C with anti-Flag beads (Sigma-Aldrich). Adar1-containing immunocomplexes were washed twice with cell lysis buffer and incubated with 500 ng RNA substrate and 300 μg nuclear extracts of BMMCs in 22 mM Tris-HCl (pH 7.5), 40 mM KCl, 10 mM NaCl, 6.5% glycerol, 0.5 mM DTT, 0.1 mM 2-mercaptoethanol, 0.05% NP-40 and RNase Inhibitor for 4 h at 30°C . The RNA was recovered by TRIzol extraction and ethanol precipitation. For the Adar1 binding assay, the beads from incubation with RNA substrate and nuclear extracts of BMMC were washed three times with ice-cold NT2 buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM MgCl₂, 0.05% NP40). The beads were resuspended by TRIzol reagent. RNA was precipitated followed by reverse transcription and qPCR analysis.

5-mC affinity pull-down coupled with LC-MS/MS analysis. BMMCs (1×10^8) were swollen for 20 min in 50 ml RSB buffer (10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.5% NP-40, 3 mM MgCl₂) and centrifuged at 2,000g for 5 min at 4°C . The pellets consisting of nuclei were lysed by 90 min incubation in two volumes of nuclear lysis buffer (420 mM NaCl, 20 mM HEPES pH 7.9, 20% v/v glycerol, 2 mM MgCl₂, 0.2 mM EDTA, 0.1% NP40, protease inhibitor and 0.5 mM DTT). After centrifugation, protein concentrations in the extracts were measured by BCA assay and stored at -80°C . Ten micrograms of 5'-biotinylated cytosine or 5-mC RNAs were immobilized on 75 μl of Dynabeads MyOne C1 (Invitrogen) by incubating for 1 h at room temperature in a total volume of 350 μl of binding buffer (1 M NaCl, 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8, and 0.05% NP-40 and RNase Inhibitor). Beads containing immobilized RNAs were then incubated with 1 mg of nuclear extracts of BMMCs in a total volume of 600 μl of protein binding buffer (50 mM Tris-HCl pH8, 150 mM NaCl, 1 mM DTT, 0.25% NP-40 and complete protease inhibitors in the presence of RNase Inhibitor) for 2 h at 4°C . Protein-complex-containing beads were washed extensively and eluted. LC-MS/MS analysis was conducted with the experimental workflows as previously reported⁶.

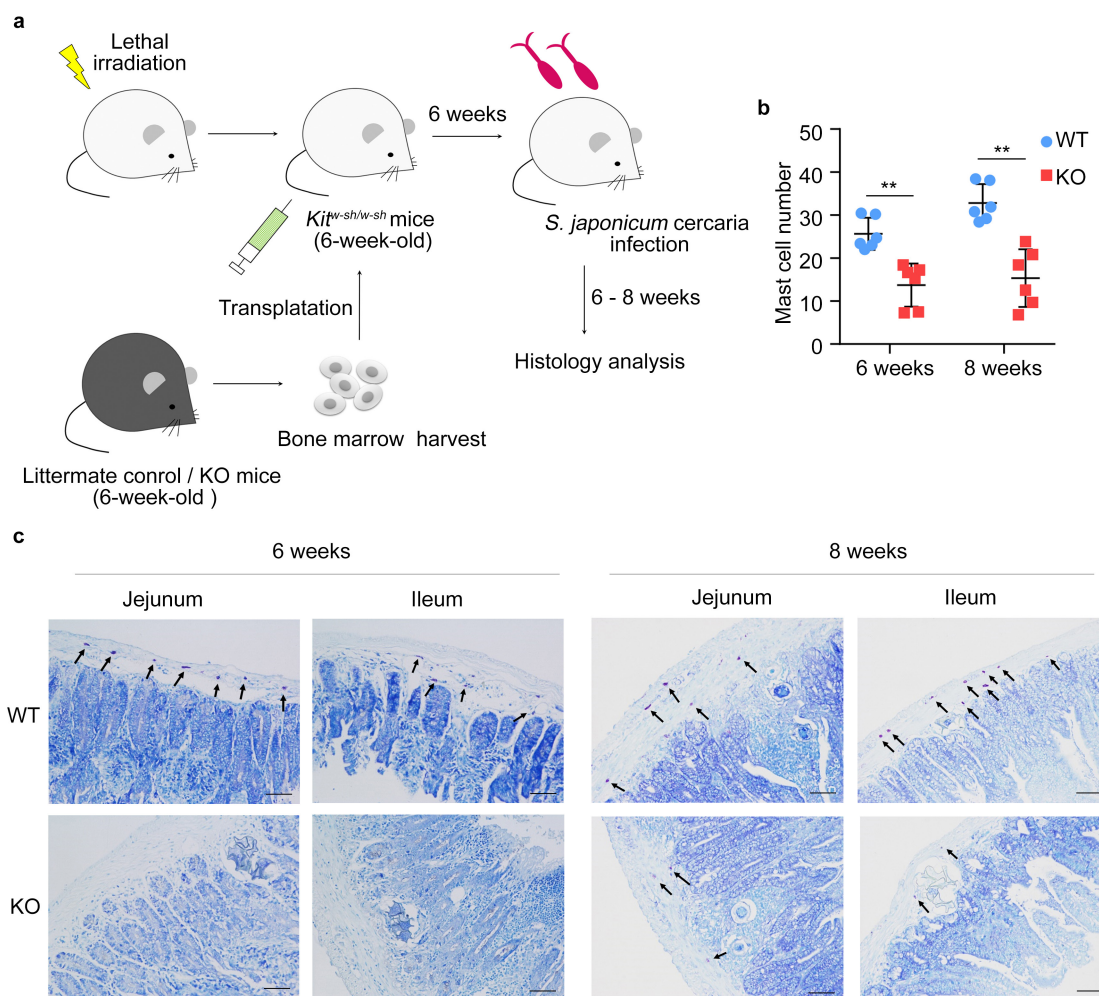
Assay of luciferase reporter gene expression. HEK293T cells were transfected with a mixture of the appropriate luciferase reporter plasmid, pRL-TK-renilla luciferase plasmid and the appropriate additional constructs using jetPEI (Polyplus). The total amount of plasmid DNA was equalized by empty control vector. Luciferase activity was measured with a Dual-Luciferase Reporter Assay System according to the manufacturer's (Promega) protocols after 24 h. Data were normalized for transfection efficiency by the division of firefly luciferase activity with that of renilla luciferase.

Statistical analysis. Error bars displayed throughout the paper represent s.e.m. or s.d. and were calculated from triplicate technical or triplicate biological

replicates described in figure legends. Sample sizes were chosen by standard methods to ensure adequate power, and no randomization of weight and sex or blinding was used for animal studies. Data shown are representative of three independent experiments, including histological images, blots and gels. No statistical method was used to predetermine sample size. Statistical significance was determined using unpaired two-sided Student's *t*-tests; **P* < 0.05; ***P* < 0.01.

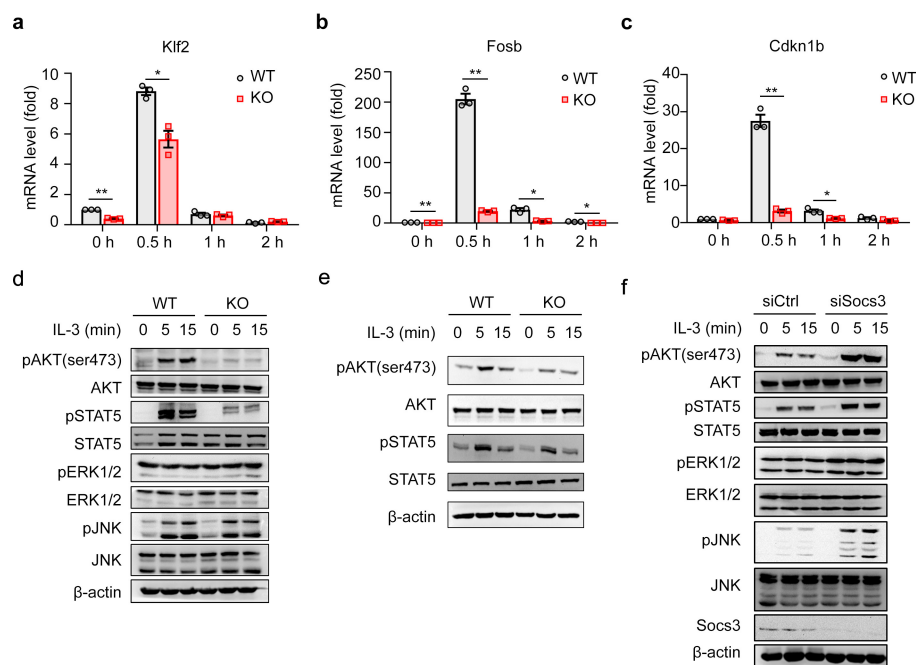
Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request. The RNA sequencing data have been deposited in NCBI Sequence Read Archive under accession numbers GSE100559, GSE100560 and GSE100719.

28. Rittirsch, D., Huber-Lang, M. S., Flierl, M. A. & Ward, P. A. Immunodesign of experimental sepsis by cecal ligation and puncture. *Nat. Protocols* **4**, 31–36 (2009).
29. Yu, M. *et al.* Tet-assisted bisulfite sequencing of 5-hydroxymethylcytosine. *Nat. Protocols* **7**, 2159–2170 (2012).
30. Li, X. *et al.* Transcriptome-wide mapping reveals reversible and dynamic N¹-methyladenosine methylome. *Nat. Chem. Biol.* **12**, 311–316 (2016).
31. Liddicoat, B. J. *et al.* RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* **349**, 1115–1120 (2015).
32. Austin, E. G. *et al.* Improvements to the HITS-CLIP protocol eliminate widespread mispriming artifacts. *BMC Genomics* **17**, 338 (2016).
33. Zhang, C. & Darnell, R. B. Mapping *in vivo* protein–RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**, 607–614 (2011).
34. Keene, J. D., Komisarow, J. M. & Friedersdorf, M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protocols* **1**, 302–307 (2006).
35. Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N. & Rechavi, G. Transcriptome-wide mapping of N⁶-methyladenosine by m⁶A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protocols* **8**, 176–189 (2013).
36. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7**, 562–578 (2012).
37. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
38. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
39. Ramaswami, G. *et al.* Identifying RNA editing sites using RNA sequencing data alone. *Nat. Methods* **10**, 128–132 (2013).
40. Peng, Z. *et al.* Comprehensive analysis of RNA-seq data reveals extensive RNA editing in a human transcriptome. *Nat. Biotechnol.* **30**, 253–260 (2012).
41. Liang, F. *et al.* BS-RNA: an efficient mapping and annotation tool for RNA bisulfite sequencing data. *Comput. Biol. Chem.* **65**, 173–177 (2016).



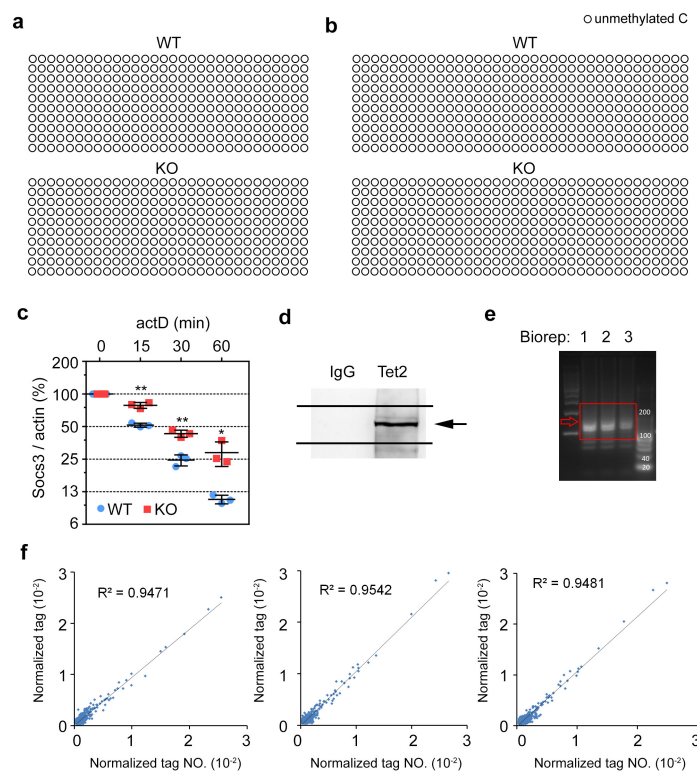
Extended Data Figure 1 | Tet2 promotes mast cell expansion during parasite infection. **a**, *In vivo* experimental design of transplantation and infection studies with bone marrow cells from Tet2-deficient (knockout) and littermate control (wild-type) mice. **b**, Quantitative assessment of toluidine blue-positive mast cells in the intestinal tissues ($n = 6$ biologically independent mice). **c**, Representative photomicrographs

of toluidine blue-stained tissue sections derived from *Kit^{W-sh/W-sh}* mice transplanted with bone marrow cells from the indicated genotypes. Arrows indicate mast cells. Scale bars, 50 μm . * $P < 0.05$, ** $P < 0.01$, unpaired two-sided Student's *t*-test. Mean and s.d. of n samples (**b**). Data are representative of three independent experiments with similar results (**c**).



Extended Data Figure 2 | Impaired IL-3 signalling pathway in Tet2-deficient myeloid cells. **a–c**, qPCR analysis of mRNA levels of indicated genes in wild-type (WT) and Tet2-deficient (KO) BMMCs treated with IL-3 (10 ng ml⁻¹). **d, e**, Immunoblot assays of the phosphorylated (p-) or total proteins in lysates of wild-type and knockout BMMCs (**d**) and bone marrow cells (**e**) stimulated with IL-3 for the indicated time. Bone marrow cells were collected from Tet2-deficient (knockout) and littermate control (wild-type) mice and pre-stimulated with IL-3 for 12 h.

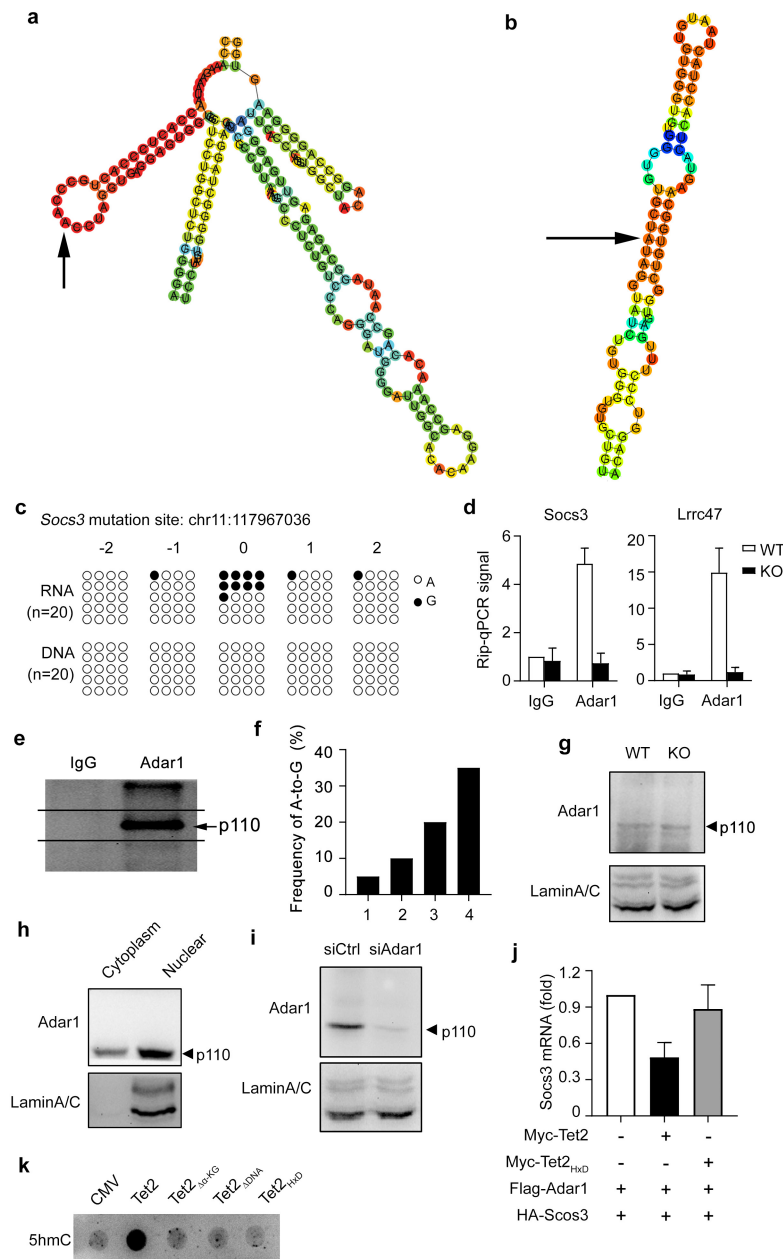
f, Immunoblot assays of the phosphorylated (p-) or total proteins in lysates of knockout BMMCs treated with non-targeting control siRNA (siCtrl) or Socs3-specific siRNA (siSocs3). Before being re-stimulated for the indicated times for subsequent analysis, BMMCs and bone marrow cells were starved for 12 h in the absence of cytokines. * $P < 0.05$, ** $P < 0.01$, unpaired two-sided Student's t -test. Mean and s.e.m. of triplicate biological replicates (**a–c**). Blots are representative of three independent experiments (**d–f**).



Extended Data Figure 3 | Tet2 binds and represses Socs3 mRNA.

a, b, Bisulfite-PCR assay of methylation states of CG dinucleotides in DNA regions of chromosome 11: 117969004-258 (**a**) or chromosome 11: 117969363-777 (**b**) in Tet2-deficient BMDCs and the control cells. **c**, Wild-type and knockout BMDCs were starved for 12 h in the absence of cytokines, and then treated with 5 mg ml⁻¹ actinomycin D (actD) for 0, 15, 30, 60 min. Socs3 mRNA decay was quantified by qPCR and normalization by β -actin. **d**, Immunoblot of Tet2 immunoprecipitation during CLIP. Black line indicates region excised for CLIP library preparation. **e**, PCR amplification products from CLIP experiments

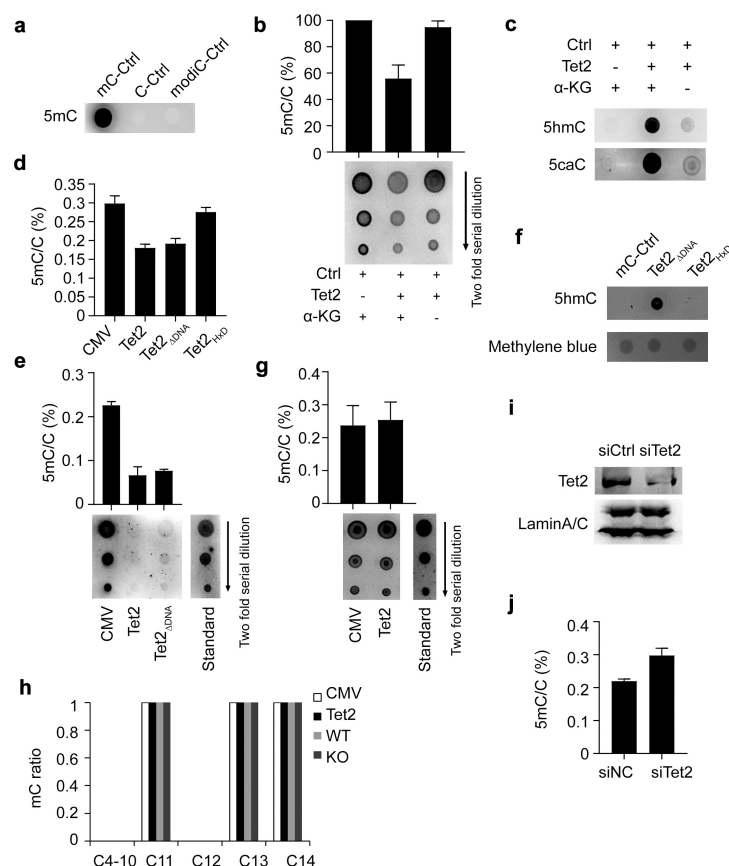
before indexing. Red box indicates gel region where DNA products were extracted for further indexing and high-throughput sequencing. Biorep 1, 2 and 3 are biological replicates from three BMDC samples which have different culture start dates and crosslinked end dates. **f**, Pairwise correlation analysis between biological replicates with normalized tag numbers in common peaks (from left to right: biorep 1 versus biorep 2, biorep 1 versus biorep 3, biorep 2 versus biorep 3). * $P < 0.05$, ** $P < 0.01$, unpaired two-sided Student's t -test. Mean and s.d. of triplicate biological replicates (**c**). Blots are representative of three independent experiments (**d**).



Extended Data Figure 4 | Adar1 binds and represses Socs3 mRNA.

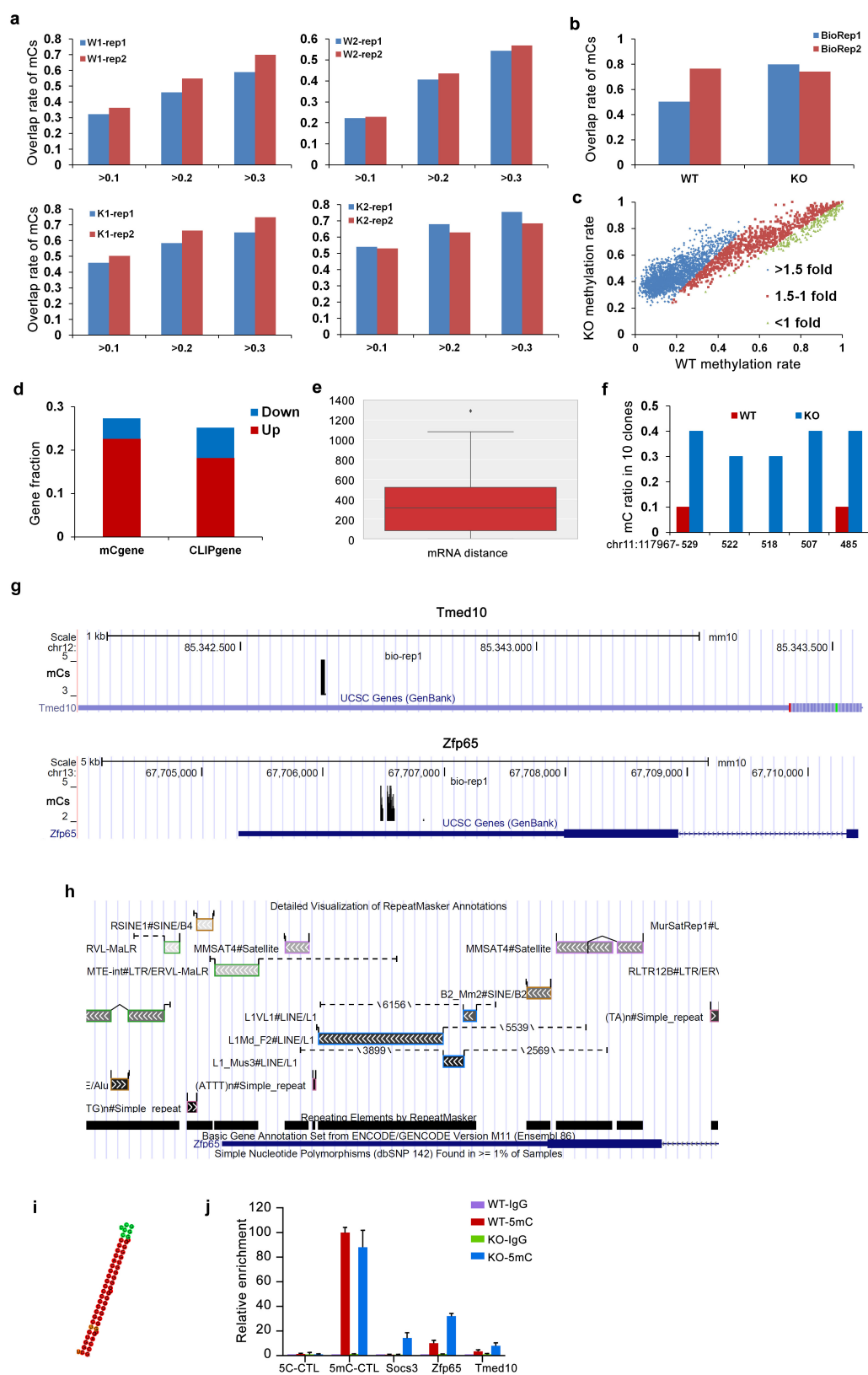
a, b, ViennaRNA prediction of secondary structure of sequences near editing sites in *Socs3* and *Lrrc47* 3' UTR. Arrows highlight edited adenosines. **c**, Experimental validation of the A-to-G mutation (0) and the nearby adenosines (-2, -1, 1, 2) in control BMMCs. **d**, RIP-qPCR analysis of *Socs3* or *Lrrc47* transcripts in RNAs from anti-Adar1 immunoprecipitated wild-type and knockout BMMCs lysates. **e**, Immunoblot of ADAR1 immunoprecipitation during CLIP. Black line indicates region excised for CLIP RNA preparation. **f**, RT-PCR sequencing assay of A-to-G mutation frequencies in *Socs3* mRNA from wild-type BMMCs at the indicated culture stages. **g**, Immunoblot of Adar1 protein

expression in BMMCs from wild-type and Tet2-deficient (knockout) mice. **h**, Immunoblot of Adar1 among cytoplasm and nuclear proteins of BMMCs. **i**, Immunoblot of Adar1 protein expression in BMMCs treated with non-targeting control siRNA (siCtrl) or Adar1-specific siRNA (siAdar1). **j**, qPCR analysis of HEK293T cells transiently transfected for 24 h with vectors coding haemagglutinin-tagged Socs3, Flag-tagged Adar1 and indicated Myc-tagged Tet2 mutants. **k**, Dot blot assays of 5-hmC levels in 10 ng DNA from Tet2- and Tet2 mutant-overexpressed HEK293T cells. Error bars, s.d. of triplicate technical replicates (**d, j**). Blots are representative of three independent experiments (**e, g-i, k**).



Extended Data Figure 5 | Tet2 promotes cytosine demethylation of mRNA. **a**, One microgram of *in vitro* transcribed RNAs containing 1% 5-mC, or 3% mixture of 5-hmC, 5-fC and 5-caC was analysed by dot blots using 5-mC antibody. **b**, **c**, The 5-mC levels in mRNAs (**b**) and 5-hmC and 5-caC levels of DNAs (**c**) from *in vitro* Tet2 oxidation assay with or without α-KG were analysed by dot blots. Twofold gradient dilutions of 20 ng synthetic *Socs3* mRNAs (**b**) and 10 ng DNAs (**c**) after oxidation were used for quantification. **d**, LC-MS for quantifying 5-mC levels of mRNAs from HEK293T cells overexpressing the indicated mutant forms of Tet2. **e**, **g**, Dot blot assays of 5-mC levels in 800 ng mRNAs (**e**) and 1 μg total RNA (**g**) from Tet2- and Tet2^{ΔDNA} mutant-overexpressed HEK293T cells. Twofold gradient dilutions of 800 ng *in vitro* transcribed *Socs3*

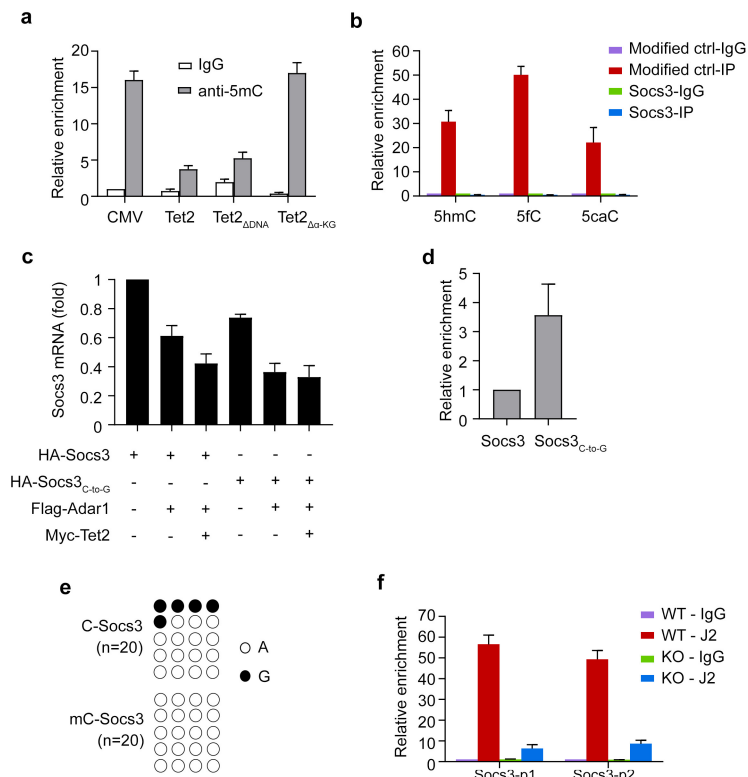
mRNAs containing 0.4% 5-mCs were used for the dilution curve of grey value-based quantification. **f**, *In vitro* RNA 5-mC oxidation assay of Tet2 mutants. The overexpressed Myc-tagged Tet2 mutants immunoprecipitated from HEK293T cells were subjected to *in vitro* oxidation. Oxidized RNAs pretreated with DNase were used for dot blot analysis of 5-hmC levels. **h**, Bisulfite-PCR assay of the 4th to 14th cytosines in tRNA^{Asp(GUC)} in Tet2-overexpressed HEK293T cells or Tet2-deficient BMMCs and the control cells. **i**, **j**, Immunoblot of Tet2 protein expression and LC-MS for quantifying 5-mC levels of mRNAs in HEK293T cells treated with non-targeting control siRNA (siCtrl) or Tet2-specific siRNA (siTet2). Mean and s.d. of triplicate technical replicates (**b**, **d**, **e**, **g**, **j**). Blots are representative of three independent experiments (**a**–**c**, **e**–**g**, **i**).



Extended Data Figure 6 | See next page for caption.

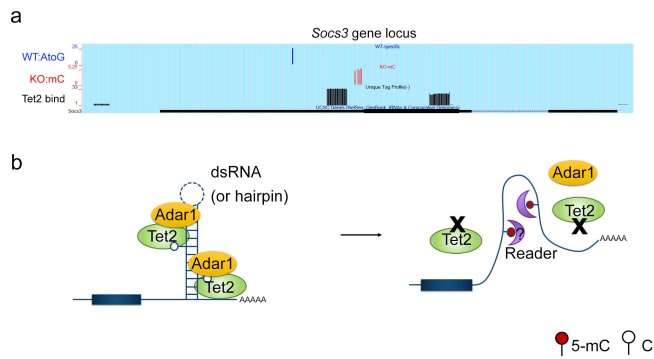
Extended Data Figure 6 | Specific profiles of mRNA 5-mCs in Tet2-deficient BMMCs. **a**, Overlap rates of methylcytosines with methylation levels above the indicated values in bisulfite sequencing assay between indicated technical replicates for Tet2-deficient (knockout, K1/2) and control (wild-type, W1/2) groups. **b**, Overlap rates of methylcytosines between the two biological replicates from common cytosines with read coverage above four. **c**, Methylcytosines in the knockout group were chosen, and mean methylation rates of these methylcytosine sites in both the wild-type and knockout groups were categorized with indicated variation folds and are presented in the scatter plot. Different colours indicate the variation of mean methylation levels of each of the methylcytosines in the knockout group compared with those in the wild-type group. **d**, Fraction of genes associated with knockout group-specific methylcytosines (mCgene) or CLIP peaks (CLIPgene) with variations of mRNA levels (>1.3 -fold, up; <0.77 -fold, down; $P < 0.05$) in the knockout group, compared with the control group. **e**, Exon-located CLIP peak and methylcytosine in the same gene were chosen, and the distance in mature mRNA between the CLIP peak boundary and the methylcytosine clusters

with the shortest gap was calculated. These distances for each of the genes are presented in the box plot (centre, median; box boundaries, 25% and 75% percentiles; whiskers, 1.5-fold interquartile range; diamond, outlier; $n = 11$ distance values). **f**, Bisulfite-PCR sequencing assay of cytosine sites with methylation-supported reads in the 3' UTR of *Socs3*. **g**, Genome browser views of gene loci containing 5-mCs in the Tet2-deficient group. Black signals indicate mean mC-supporting read numbers of all the replicates in the Tet2-deficient group. **h**, Genome browser view of the indicated region with RepeatMasker Viz containing the editing site in the 3' UTR of *Zfp65*. **i**, ViennaRNA prediction of secondary structure of sequences containing the methylation sites in the 3' UTR of *Tmed10*. **j**, qPCR analysis of gene transcripts from anti-5-mC immuno-selected RNAs from total RNAs of wild-type and knockout BMMCs. Unmethylated and methylated spike-ins as the negative and positive controls. Cytosine with coverage above 4, at least two reads supporting methylation and methylation level equal or above 0.1 was chosen as methylcytosine, considering both bioinformatic and biological significance. Mean and s.d of triplicate technical replicates (j).



Extended Data Figure 7 | Cytosine methylation in the 3' UTR of *Socs3* inhibits dsRNA structure. **a**, qPCR analysis of overexpressed *Socs3* transcripts from anti-5-mC immuno-selected RNAs from total RNAs of HEK293T cells transfected with Tet2 or Tet2 mutants. **b**, qPCR analysis of *Socs3* transcripts from specific-modification antibodies immuno-selected RNAs from total RNAs of wild-type BMMCs. Unmodified and modified spike-ins as the negative and positive controls. **c**, qPCR analysis of HEK293T cells transiently transfected for 24 h with vectors coding haemagglutinin-tagged wild-type *Socs3* or C-to-G mutant *Socs3*

(*Socs3*_{C-to-G}), with or without Flag-tagged Adar1 and Myc-tagged Tet2. **d**, RIP-qPCR analysis of *Socs3* 3' UTR levels in RNAs from Flag-tagged Adar1-immunoprecipitated HEK293T cell lysates overexpressed with *Socs3* or *Socs3*_{C-to-G} together with Adar1. Lysates (1%) were used for normalization as input. **e**, A-to-I editing rates in *Socs3* 3' UTR with cytosine or 5-mC after Adar1 editing *in vitro*. **f**, *Socs3* transcript levels determined by RT-qPCR from J2 immuno-selected dsRNA; p1, primer 1; p2, primer 2. Mean and s.d. of triplicate technical replicates (**a–d, f**). Data are representative of three independent experiments (**e**).



Extended Data Figure 8 | Schematic illustration of Tet2-mediated repression of *Socs3* via Adar1. **a**, Genome browser view of sequencing data on *Socs3* locus. Blue, A-to-G mutant reads in wild-type BMMCs; red, mean mC-supporting read numbers in knockout BMMCs; black, CLIP tag coverage. **b**, Tet2 promotes mRNA cytosine demethylation for effective formation of dsRNA which is bound by Adar1, leading to the suppression of *Socs3* expression at the post-transcriptional level.

Extended Data Table 1 | Sequences of PCR primers used in this study

Name	Prime	Sequence
<i>mSocs3_CLIP</i>	Forward	5' - GCGCTTTGATTTGGTTTGAT - 3'
	Reverse	5' - GGTTATTTCTTTGGCCAGCA - 3'
<i>mLrrc47_CLIP/RIP</i>	Forward	5' - CTGACAGGCTCCTGTAGGTGT - 3'
	Reverse	5' - TACAGCACACCCACAGATACCTAT - 3'
<i>mSocs3_RIP</i>	Forward	5' - CCTTTGACAAGCGGACTCTC - 3'
	Reverse	5' - GCCAGCATAAAAACCCTTCA - 3'
<i>mSocs3_P1</i>	Forward	5' - TATTCTGGGGGCGAGAAGAT - 3'
	Reverse	5' - ATCCAGGAACTCCCGAATG - 3'
<i>mSocs3_P2</i>	Forward	5' - ACATGGCACAAGCACAAAAA - 3'
	Reverse	5' - GCTGGCACTTGAAAGAA - 3'
<i>mAdar1</i>	Forward	5' - TGAGCATAGCAAGTGGAGATACC - 3'
	Reverse	5' - GCCGCCCTTTGAGAACTCT - 3'
<i>unmodified-Ctrl</i>	Forward	5' - ATTGTATGTATTGGTTTATTG - 3'
	Reverse	5' - TTATCACATTCAAACATTAAT - 3'
<i>modified-Ctrl</i>	Forward	5' - TAGATAGTAAATATAATGTGAGA - 3'
	Reverse	5' - ATAAATCATCAACAAAACACAA - 3'
<i>mZfp65</i>	Forward	5' - GTGTGGAGACTTTGCCATT - 3'
	Reverse	5' - AAATGGTGTCAGCGTTTGGT - 3'
<i>mTmed10</i>	Forward	5' - CCAGGTAGAGTAGTCCATCCC - 3'
	Reverse	5' - AGGTTACACTCTAGATGACCCA - 3'
<i>mKlf2</i>	Forward	5' - CTCAGCGAGCCTATCTTGCC - 3'
	Reverse	5' - CACGTTGTTTAGGTCCTCATCC - 3'
<i>mFosb</i>	Forward	5' - TTTTCCCGGAGACTACGACTC - 3'
	Reverse	5' - GTGATTGCGGTGACCGTTG - 3'
<i>mCdkn1b</i>	Forward	5' - TCAAACGTGAGAGTGCTAACG - 3'
	Reverse	5' - CCGGGCCGAAGAGATTTCTG - 3'
<i>mβ-actin</i>	Forward	5' - AGTGTGACGTTGACATCCGT - 3'
	Reverse	5' - GCAGCTCAGTAACAGTCCGC - 3'
<i>mSocs3_editing</i>	Forward	5' - GACCTCTCTCCTCCAACGTG - 3'
	Reverse	5' - TGCAAAGTCTGAGTTGAACTGG - 3'
<i>mLrrc47_editing</i>	Forward	5' - CTGACAGGCTCCTGTAGGTGT - 3'
	Reverse	5' - GCATACCCACACCCAGATAC - 3'
<i>Socs3_bisulfite PCR</i>	Forward	5' - GTTTTAAGATTTTGTATTTTAAAG - 3'
	Reverse	5' - AATACACCAACTTAAATACACAATC - 3'
<i>Socs3pro1_bisulfite PCR</i>	Forward	5' - AGAGTAGTGATTAAATATTATAAGAAGAT - 3'
	Reverse	5' - TAAATCTACAAAAAACTCCCC - 3'
<i>Socs3pro2_bisulfite PCR</i>	Forward	5' - TTTGGATTTGTTTATAGGTAAATGT - 3'
	Reverse	5' - CTTAAACTAAAACCTCCAAAACCC - 3'

Mitochondrial translation requires folate-dependent tRNA methylation

Raphael J. Morscher^{1,2}, Gregory S. Ducker^{1,2}, Sophia Hsin-Jung Li³, Johannes A. Mayer⁴, Zemer Gitai³, Wolfgang Sperl⁴ & Joshua D. Rabinowitz^{1,2}

Folates enable the activation and transfer of one-carbon units for the biosynthesis of purines, thymidine and methionine^{1–3}. Antifolates are important immunosuppressive⁴ and anticancer agents⁵. In proliferating lymphocytes⁶ and human cancers^{7,8}, mitochondrial folate enzymes are particularly strongly upregulated. This in part reflects the need for mitochondria to generate one-carbon units and export them to the cytosol for anabolic metabolism^{2,9}. The full range of uses of folate-bound one-carbon units in the mitochondrial compartment itself, however, has not been thoroughly explored. Here we show that loss of the catalytic activity of the mitochondrial folate enzyme serine hydroxymethyltransferase 2 (SHMT2), but not of other folate enzymes, leads to defective oxidative phosphorylation in human cells due to impaired mitochondrial translation. We find that SHMT2, presumably by generating mitochondrial 5,10-methylenetetrahydrofolate, provides methyl donors to produce the taurinomethyluridine base at the wobble position of select mitochondrial tRNAs. Mitochondrial ribosome profiling in SHMT2-knockout human cells reveals that the lack of this modified base causes defective translation, with preferential mitochondrial ribosome stalling at certain lysine (AAG) and leucine (UUG) codons. This results in the impaired expression of respiratory chain enzymes. Stalling at these specific codons also occurs in certain inborn errors of mitochondrial metabolism. Disruption of whole-cell folate metabolism, by either folate deficiency or antifolate treatment, also impairs the respiratory chain. In summary, mammalian mitochondria use folate-bound one-carbon units to methylate tRNA, and this modification is required for mitochondrial translation and thus oxidative phosphorylation.

The major source of folate one-carbon (1C) units in mammalian cells is the amino acid serine^{1–3}. Transfer of the 1C unit of serine to tetrahydrofolate (THF) can occur in either the cytosol or the mitochondrion, via the enzyme SHMT1 or SHMT2, respectively¹⁰ (Fig. 1a). Evidence from stable isotope tracing indicates that cancer cells predominantly use SHMT2 to catabolize serine, exporting the resulting 1C units to the cytosol to support nucleotide synthesis^{2,9}. The extent to which 1C unit production via SHMT2 is also important to support mitochondrial health has yet to be determined.

When characterizing a set of human HCT116 colon cancer CRISPR-deletion cell lines that lack folate 1C enzymes, we discovered that the loss of SHMT2 induces a change in media colour indicative of enhanced extracellular acidification (Extended Data Fig. 1a). Quantitative analysis of the media confirmed increased glucose uptake and lactate secretion. This effect was specific to SHMT2: loss of other core 1C enzymes, including SHMT1 and mitochondrial enzymes such as MTHFD2 and MTHFD1L, did not induce glycolysis (Fig. 1b, Extended Data Fig. 1b).

A common cause of increased glycolytic flux is respiratory deficiency¹¹. Loss of SHMT2 reduced both basal respiration and

maximal respiratory capacity and decreased the NAD⁺/NADH ratio in several HCT116 (Fig. 1c, Extended Data Fig. 1c) and HEK293T SHMT2-knockout clones. Knockout cell lines lacking other core folate enzymes did not show impaired respiration (Extended Data Fig. 1d, e). Consistent with respiratory chain deficiency, the loss of SHMT2 decreased glucose flux into the tricarboxylic acid (TCA) cycle intermediate citrate, with an increased fraction of citrate instead being produced by reductive carboxylation¹². As reported recently in other models of mitochondrial damage^{13,14}, the pool size of TCA cycle metabolites and associated amino acids was also decreased (Extended Data Fig. 1f, g). To identify the cause of respiratory deficiency, we examined the abundances of several mitochondrial proteins, and found decreased abundances of complex I, IV and V subunits with retained levels of complexes II and III and markers of mitochondrial mass (Fig. 1d and Extended Data Fig. 1h, i). Thus, SHMT2 is required to maintain the levels of several mitochondrial respiratory chain proteins.

Given that the loss of SHMT2, but not of the immediate downstream enzymes of mitochondrial 1C metabolism, caused impaired oxidative phosphorylation, we wondered whether the phenotype reflected a requirement for the catalytic activity of SHMT2, or alternatively a non-catalytic role of SHMT2, perhaps related to its reported interaction with the mitochondrial nucleoid¹⁵. Accordingly, in the SHMT2-knockout background, we stably re-expressed catalytically inactive SHMT2 (p.Glu98Leu/p.Tyr106Phe), pyridoxal 5'-phosphate (PLP)-binding mutant SHMT2 (p.Lys280Gln) or wild-type SHMT2 protein (Extended Data Figs 2, 3a). Re-expression of wild-type protein, but not the catalytically inactive mutants, rescued the oxidative phosphorylation deficiency (Fig. 1e and Extended Data Fig. 3b) and normalized glycolytic flux (Extended Data Fig. 3c). In addition, SHMT2 re-expression rescued the growth defect of SHMT2-knockout cells⁹ (Extended Data Fig. 3d) and normalized 1C metabolism (Extended Data Fig. 3e, f). Thus, mitochondrial SHMT catalytic activity is crucial to sustain oxidative phosphorylation.

Two compartment-specific uses of mitochondrial folate 1C units have been reported: the local biosynthesis of deoxythymidine triphosphate (dTTP)^{16,17} and of *N*-formylmethionine (f-Met)^{18,19} (Fig. 2a). The production of dTTP requires 5,10-methylene-THF (methylene-THF), whereas f-Met requires 10-formyl-THF (formyl-THF). SHMT2 is upstream of both compounds. By contrast, MTHFD2 sits between methylene-THF and formyl-THF. The lack of an oxidative phosphorylation phenotype with MTHFD2 knockout led us to hypothesize that methylene-THF is the required 1C species. Consistent with this, SHMT2-knockout cell lines showed unchanged N-terminal f-Met levels of the mitochondrially translated COX1 peptide (encoded by the *MT-CO1* gene)¹⁹ (Extended Data Fig. 4a). To confirm that methylene-THF is the required species, we generated SHMT2/MTHFD2 double-deletion cells and supplemented them with methylglycine (sarcosine), which can produce mitochondrial methylene-THF

¹Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey 08544, USA. ²Department of Chemistry, Princeton University, Princeton, New Jersey 08544, USA.

³Department of Molecular Biology, Princeton University, Princeton, New Jersey 08544, USA. ⁴Department of Pediatrics, Salzburger Landeskliniken and Paracelsus Medical University, Salzburg 5020, Austria.

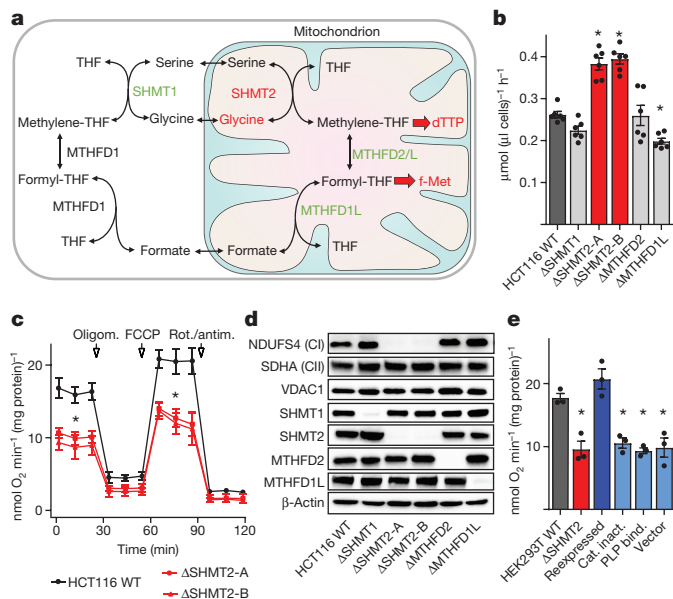


Figure 1 | Mitochondrial respiratory chain function is dependent on SHMT2 catalytic activity. **a**, 1C pathway and known mitochondrial products. **b**, Lactate secretion of HCT116-knockout cell lines ($n=6$). ΔSHMT2-A and ΔSHMT2-B denote two separate SHMT2-knockout lines. WT, wild type. **c**, Oxygen consumption rate measured by Seahorse XF analyser ($n=3$). FCCP denotes a mitochondrial uncoupling agent. Oligom., oligomycin; Rot./antim., rotenone/antimycin. **d**, Immunoblot for mitochondrial respiratory complex I and II (CI and CII) proteins (NDUF54 and SDHA, respectively), 1C enzymes, and a marker of mitochondrial mass (VDAC1). **e**, Basal respiration ($n=3$) upon re-expression of wild-type or catalytically deficient mutant forms of SHMT2 in HEK293T knockout cell lines. Data are mean \pm s.e.m. n indicates the number of biological replicates, which for the Seahorse experiments refers to independent plates on separate days. * $P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values). Cat. inact., catalytically inactive SHMT2; PLP bind., PLP binding-deficient SHMT2.

via sarcosine dehydrogenase. Although sarcosine is not a preferred 1C source and its feeding was insufficient to restore oxidative phosphorylation in SHMT2 single gene knockout cell lines, it fully restored oxidative phosphorylation in the SHMT2/MTHFD2 double-knockout cells,

in which drainage of methylene-THF to formyl-THF is blocked (Fig. 2b and Extended Data Fig. 4b). Thus, mitochondrial methylene-THF is required to maintain respiratory capacity.

Depletion of mitochondrial dTTP and/or accumulation of uridine nucleotides have been shown to induce mitochondrial respiratory chain deficiency by promoting mitochondrial DNA damage^{17,20}. However, SHMT2-knockout cells showed no evidence of altered mitochondrial DNA copy number (Fig. 2c and Extended Data Fig. 4c), deletions (Extended Data Fig. 4d), or mutations (Extended Data Fig. 4g, h and Supplementary Table 1). Moreover, whole-cell RNA sequencing revealed normal transcript levels for both nuclear and mitochondrial-encoded respiratory chain protein subunits (Fig. 2d, Extended Data Fig. 4e, f). Thus, the dependence of oxidative phosphorylation on SHMT2 reflects a requirement for mitochondrial methylene-THF for a purpose other than supplying local dTTP to maintain mitochondrial DNA.

Whereas the vast majority of the approximately 1,100 mitochondrial proteins are imported from the cytosol, 13 essential respiratory chain subunits are locally transcribed and translated²¹. These include components of complexes I, III, IV and V, but not complex II. On the basis of the normal mitochondrial transcript abundances and complex II protein levels, we hypothesized that mitochondrial methylene-THF is required for local translation. Indeed, [³⁵S]methionine incorporation assays showed decreased synthesis of certain complex I and IV subunits (Extended Data Fig. 5a). To probe mitochondrial translation further, we developed a protocol for mitochondrial ribosome profiling based on digesting unprotected mRNA with micrococcal nuclease, enriching the 55S mitochondrial ribosome, and sequencing the protected footprints (Fig. 3a and Extended Data Fig. 5b). This approach achieved more than 90% average mitochondrial transcript sequence coverage, with an average depth of at least 80 reads per codon (Supplementary Table 3 and Extended Data Fig. 5c). In SHMT2-knockout cells, the distribution of ribosome-protected footprints showed pronounced stalling at defined codon positions (Fig. 3b and Extended Data Fig. 6a). This resulted in relatively fewer actively translating ribosomes (that is, bound and not stalled) for certain subunits of respiratory chain complexes I, IV and V (Extended Data Fig. 6b). Consistent with the ribosome profiling data, enzymatic assays revealed decreased activity of complexes I, IV and V (Extended Data Fig. 6c).

We next aimed to determine the cause of ribosomal stalling. The aminoacyl-tRNA acceptor-site (A-site) coordinates of stalled ribosomes revealed notable ribosome accumulation in SHMT2-knockout cells at particular lysine and leucine codons: Lys^{AAG} and Leu^{UUG} (Fig. 3c).

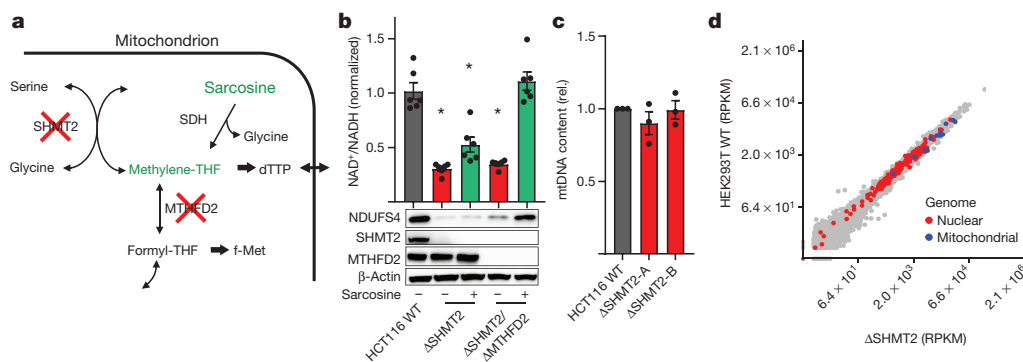


Figure 2 | SHMT2-knockout-induced respiratory chain deficiency is caused by mitochondrial methylene-THF depletion but is unrelated to dTTP synthesis. **a**, Sarcosine serves as an SHMT2-independent source of mitochondrial methylene-THF. **b**, NAD⁺/NADH ratio ($n=6$) and NDUF54 (complex I) protein expression upon sarcosine supplementation (1 mM) in SHMT2 single-knockout (ΔSHMT2) and SHMT2/MTHFD2 double-knockout (ΔSHMT2/ΔMTHFD2) cell lines compared to wild-type cells. **c**, **d**, Functional readouts for mitochondrial dTTP status based on mitochondrial DNA (mtDNA) levels ($n=3$) determined by quantitative PCR (qPCR; **c**) and gene expression determined by

RNA-seq (**d**) in SHMT2-knockout and wild-type HEK293T cells. RPKM, reads per kilobase per million mapped reads. In **d**, each data point represents the mean gene expression of two biological replicates of two independent knockout clones ($n=4$) and two wild-type replicates ($n=2$). Genes linked to OXPHOS function³⁷ are highlighted in red (nuclear-encoded) or blue (mitochondrial-encoded). Significantly differentially expressed genes are shown in Supplementary Table 2. Data are mean \pm s.e.m. n indicates the number of independent biological replicates. * $P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values).

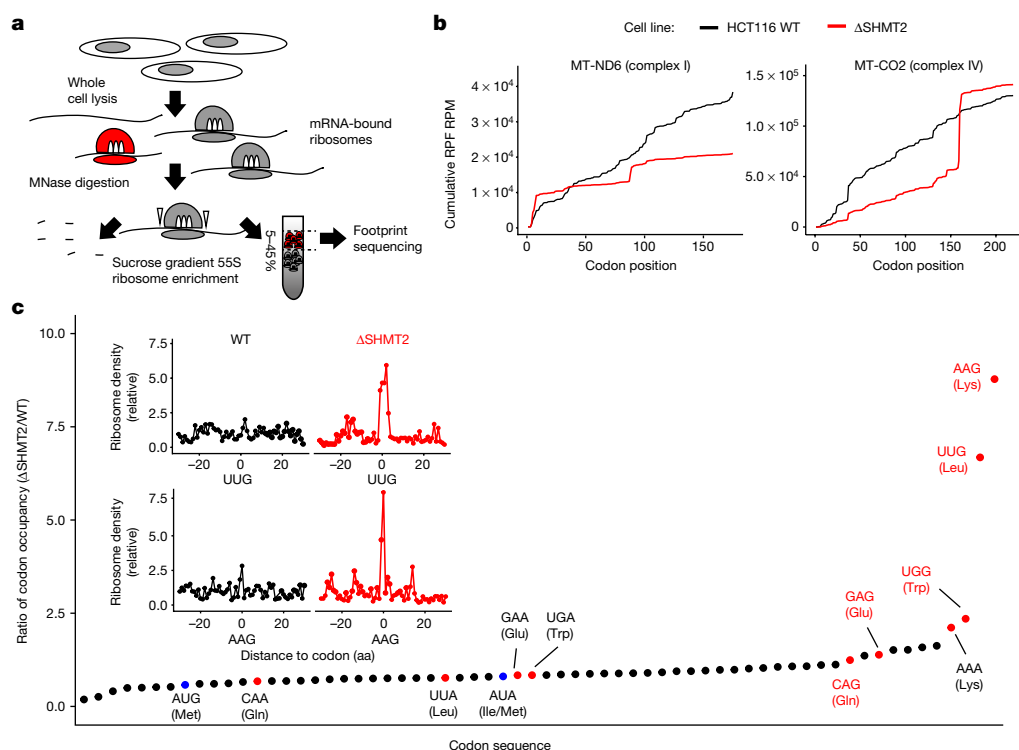


Figure 3 | Mitochondrial ribosome profiling reveals that SHMT2-knockout cells are deficient in translating specific guanosine-ending codons. **a**, Workflow of mitochondrial ribosome profiling. Translation was halted using chloramphenicol and immersion into liquid nitrogen, cells were lysed and RNA was digested using micrococcal nuclease (MNase). After sucrose gradient enrichment for mitochondrial ribosomes (shaded in red), protected fragments were sequenced. **b**, Mean cumulative ribosome density along selected mitochondrial transcripts. Additional transcripts are given in Extended Data Fig. 6a. RPF, ribosome-protected

fragment; RPM, reads per million. **c**, Mean codon-specific mitochondrial ribosome occupancy in HCT116 cells (ΔSHMT2/wild type). Red data points correspond to codons that are decoded by tRNAs carrying the 5-taurinomethyluridine (τ^m U) modification. Red labels correspond to the subset of codons that end in guanosine and thus require wobble-base pairing. Methionine codons are highlighted in blue and show no increased codon occupancy. The insert shows mean normalized ribosome density relative to UUG and AAG codon position. aa, amino acids. Data in **b** and **c** represent two technical replicates of two independent samples.

This did not seem to reflect a shortage of these amino acids or tRNAs in mitochondria, as much less stalling was observed at the corresponding Lys^{AAA} and Leu^{UUA} codons. Instead, it appeared to relate to difficulty in reading the 3' codon guanine of certain codons: those in which the 3' position identity (purine versus pyrimidine) determines the encoded amino acid ('split codon boxes'; Extended Data Fig. 6d). Increased codon occupancy was observed also for Trp^{UGG}, Glu^{GAG} and Gln^{CAG}, but not the corresponding codons with 3' adenosine. The sole exception was Met^{AUG/AUA}, in which no stalling was observed. Decoding of A/G-ending codons in split codon boxes is facilitated by methyl-derivative base modifications of the tRNA anticodon 5' nucleotide, allowing non-Watson-Crick base-pairing with the codon 3' base²² (Fig. 4a). The mitochondrial tRNA^{Met} anticodon has a 5' cytidine that is formylated with the 1C unit derived from S-adenosyl-methionine (SAM)²³. By contrast, the mitochondrial tRNAs for Lys, Leu1 (one of two mitochondrial leucine tRNAs), Trp, Glu and Gln have uridine at the 5' anticodon position²⁴. Mammalian cytosolic tRNAs with uridine at the anticodon 5' position are modified to produce 5-methoxycarbonylmethyluridine (mcm⁵U), with the methoxy carbon derived from SAM^{25,26}. Corresponding mitochondrial tRNAs are 5-taurinomethyl modified at the 5' anticodon uridine (τ^m U at position 34 of the tRNA)^{27–29}. As SHMT2/MTHFD2 double-knockout cell lines supplemented with sarcosine showed rescue of codon-specific stalling on mitochondrial ribosome profiling (Extended Data Fig. 6e), we hypothesized that the τ^m U modification is dependent on mitochondrial methylene-THF and that the observed oxidative phosphorylation defect is a consequence of impaired translation due to defective tRNA modification.

To test this hypothesis, we established a liquid chromatography–mass spectrometry (LC–MS) method for detection of modified

mitochondrial tRNA bases. In SHMT2-knockout cell lines, levels of formylcytidine were unchanged (Extended Data Fig. 7a). By contrast, τ^m U and its 2-thio derivative (in which the uracil 2-position oxygen is replaced with sulfur) were depleted to undetectable levels (Fig. 4b and Extended Data Fig. 7b). The depletion was not due to a lack of taurine, the cellular levels of which tended to be higher (Extended Data Fig. 7c). The loss of τ^m U and its 2-thio derivative was reversed upon re-expression of wild-type SHMT2 and also upon sarcosine supplementation of SHMT2/MTHFD2 double-knockout cells (Extended Data Fig. 7d, e). Thus, SHMT2 or an alternative source of mitochondrial methylene-THF is required to produce τ^m U.

To verify that the methyl group of τ^m U is coming from methylene-THF made by SHMT2, we conducted stable isotope tracing with ¹³C-labelled serine (the SHMT2 substrate) versus [¹³C]methionine (which feeds into SAM). As expected, methionine labelled formylcytidine. By contrast, serine labelled τ^m U, but methionine did not (Extended Data Fig. 7f). Thus, SHMT2 is required for oxidative phosphorylation because it produces mitochondrial methylene-THF for tRNA taurinomethylation. To our knowledge, this is the first direct evidence of folate-dependent macromolecule modification in mammalian cells.

The enzyme complex that catalyses the τ^m U base modification in mitochondria comprises two proteins, MTO1 and GTPBP3, both of which are encoded by genes that are mutated in human mitochondrial diseases^{30,31}. Orthologues of the enzyme complex, forming the prokaryotic tRNA-modifying MnmE/GidA complex, have been shown to use THF-bound 1C units³². We therefore explored whether the mitochondrial translation defect upon SHMT2 loss matches human diseases proposed to affect taurinomethylation of mitochondrial tRNAs^{27,28,30,31}. Indeed, MTO1 knockout in HCT116 cells caused loss

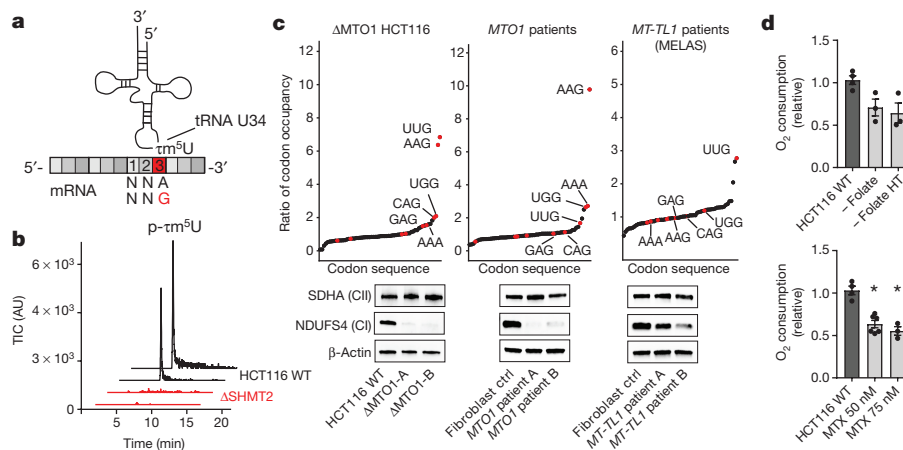


Figure 4 | MTO1/GTPBP3-dependent tRNA methylation requires mitochondrial methylene-THF. **a**, Interaction of tRNA position 34 anticodon loop modified base with mRNA codon 3 position A/G, forming a non-Watson–Crick base pair. **b**, Total ion chromatogram of 5-taurinomethyluridine monophosphate (p- τm^5U) ($m/z = 460.043$) from digested mitochondrial tRNAs. 5-formylcytidine monophosphate was not altered (Extended Data Fig. 7a). **c**, Mean codon-specific mitochondrial ribosome occupancy for MTO1-knockout ($\Delta MTO1$) HCT116 cell lines and primary patient-derived fibroblasts carrying *MTO1* mutations

of τm^5U (Extended Data Fig. 7e). Mitochondrial ribosome profiling of these engineered MTO1-deletion cells showed an increase in codon occupancy at the same codons as with SHMT2 knockout (Fig. 4c). In both cases, the strongest stalling occurred at Lys^{AAG} and Leu^{UUG}, potentially because these codons have, in addition to the wobble base pairing, only A–U base pairings versus the stronger C–G base pairings in Glu^{GAG}, Gln^{CUG} and Trp^{UGG} codons³³. Mitochondrial ribosome profiling of primary fibroblasts of two patients carrying different *MTO1* missense mutations revealed increased codon occupancy selectively at Lys^{AAG}, but not Leu^{UUG}. Thus, these hypomorphic mutations seem to cause a selective defect in taurinomethylation of the Lys tRNA. Of all adenosine-ending codons, Lys^{AAA} showed the greatest increase in codon occupancy across all cell lines, being most pronounced in the *MTO1* patients (Fig. 4c).

Notably, in both SHMT2 and MTO1 deletion cells, stalling did not affect all AAG and UUG codons uniformly, but occurred most strongly at the same specific gene locations (Extended Data Fig. 7g). Mapping of AAG and UUG codons at stalling sites relative to mRNA secondary structure³⁴ did not reveal any clear pattern (Extended Data Fig. 8a, b and Supplementary Tables 4, 5). Mapping onto the structure of the protein being synthesized (Extended Data Fig. 8c), however, showed a trend towards stalling at transitions between transmembrane helices and non-membrane domains (Supplementary Table 5). Thus, stalling due to the defective codon–anticodon interaction might be exacerbated by particular protein sequence features.

Defined mutations in the Lys and Leu1 mitochondrial tRNAs (which decode the most strongly affected Lys^{AAG} and Leu^{UUG} codons) result in a tRNA-specific τm^5U modification defect, causing the human mitochondrial disorders MERRF and MELAS^{27,28}. Mitochondrial ribosome profiling of fibroblasts from two patients with MELAS due to the m.3243A>G mutation in the *MT-TL1* gene revealed, as expected³⁵, increased occupancy at Leu^{UUG} but not Lys^{AAG} or Leu^{UUA} (Fig. 4d). For unknown reasons, increased ribosome occupancy for either Ser^{AGU} or Thr^{ACG} was also observed in individual patients (Extended Data Fig. 9a). The extent of stalling and complex I depletion was less than with nuclear *SHMT2* or *MTO1* mutations, presumably owing to the heteroplasmic nature of the mitochondrial tRNA mutation (Fig. 4c and Extended Data Fig. 9b). Collectively these observations highlight a common biochemical mechanism that links mitochondrial folate metabolism with a considerable fraction of inborn errors of mitochondrial metabolism^{27,28,30,31}.

or the *MT-TL1* m.3243A>G MELAS variant ($n = 2$). Corresponding immunoblots are shown below. Individual patient data are in Extended Data Fig. 9a. **d**, Basal respiration rates measured using the Seahorse XF analyser. Data were collected after growth in the absence (–) of folate for 5 passages or in the presence of the indicated methotrexate (MTX) concentration for 96 h ($n = 3$, except HCT116 WT $n = 4$ and MTX 50 nM $n = 6$). HT, 100 μ M hypoxanthine and 16 μ M thymidine. Data are mean \pm s.e.m. n indicates the number of biological replicates. * $P < 0.01$, two-tailed Student's *t*-test (see Supplementary Table 7 for exact *P* values).

On the basis of these observations, we wondered whether folate deficiency, which in pregnancy causes neural tube defects^{1–3}, could also result in mitochondrial impairment. Consistent with such a possibility, growth of HCT116 wild-type cells in folate-deficient media resulted in decreased levels of complex I enzymes and a reduced basal respiration rate (Fig. 4d and Extended Data Fig. 10a), without affecting mitochondrial DNA content (Extended Data Fig. 10b). In individuals with adequate nutrition, functional folate deficiency can be caused by antifolate therapy for autoimmunity or cancer. Low doses of the antifolate methotrexate, resulting in up to 100 nM circulating drug concentrations, are commonly used to treat autoimmune diseases such as rheumatoid arthritis^{4,36}. In cell culture, nanomolar concentrations of methotrexate resulted in decreased levels of complex I enzymes and oxygen consumption (Fig. 4d and Extended Data Fig. 10c). Thus, the essential role for folate metabolism in mitochondrial translation may contribute to the clinical manifestations of folate deficiency and the clinical efficacy of antifolate therapies.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 June; accepted 14 December 2017.

Published online 24 January 2018.

1. Fox, J. T. & Stover, P. J. Folate-mediated one-carbon metabolism. *Vitam. Horm.* **79**, 1–44 (2008).
2. Tibbetts, A. S. & Appling, D. R. Compartmentalization of mammalian folate-mediated one-carbon metabolism. *Annu. Rev. Nutr.* **30**, 57–81 (2010).
3. Ducker, G. S. & Rabinowitz, J. D. One-carbon metabolism in health and disease. *Cell Metab.* **25**, 27–42 (2017).
4. Lipsky, P. E. *et al.* Infliximab and methotrexate in the treatment of rheumatoid arthritis. *N. Engl. J. Med.* **343**, 1594–1602 (2000).
5. Chabner, B. A. & Roberts, T. G. Jr. Timeline: chemotherapy and the war on cancer. *Nat. Rev. Cancer* **5**, 65–72 (2005).
6. Ron-Harel, N. *et al.* Mitochondrial biogenesis and proteome remodeling promote one-carbon metabolism for T cell activation. *Cell Metab.* **24**, 104–117 (2016).
7. Nilsson, R. *et al.* Metabolic enzyme expression highlights a key role for MTHFD2 and the mitochondrial folate pathway in cancer. *Nat. Commun.* **5**, 3128 (2014).
8. Kim, D. *et al.* SHMT2 drives glioma cell survival in ischaemia but imposes a dependence on glycine clearance. *Nature* **520**, 363–367 (2015).
9. Ducker, G. S. *et al.* Reversal of cytosolic one-carbon flux compensates for loss of the mitochondrial folate pathway. *Cell Metab.* **23**, 1140–1153 (2016).
10. Garrow, T. A. *et al.* Cloning of human cDNAs encoding mitochondrial and cytosolic serine hydroxymethyltransferases and chromosomal localization. *J. Biol. Chem.* **268**, 11910–11916 (1993).

11. Gohil, V. M. *et al.* Nutrient-sensitized screening for drugs that shift energy metabolism from mitochondrial respiration to glycolysis. *Nat. Biotechnol.* **28**, 249–255 (2010).
12. Mullen, A. R. *et al.* Reductive carboxylation supports growth in tumour cells with defective mitochondria. *Nature* **481**, 385–388 (2011).
13. Sullivan, L. B. *et al.* Supporting aspartate biosynthesis is an essential function of respiration in proliferating cells. *Cell* **162**, 552–563 (2015).
14. Birsoy, K. *et al.* An essential role of the mitochondrial electron transport chain in cell proliferation is to enable aspartate synthesis. *Cell* **162**, 540–551 (2015).
15. Iborra, F. J., Kimura, H. & Cook, P. R. The functional organization of mitochondrial genomes in human cells. *BMC Biol.* **2**, 9 (2004).
16. Brown, S. S., Neal, G. E. & Williams, D. C. Subcellular distribution of some folic acid-linked enzymes in rat liver. *Biochem. J.* **97**, 34C–36C (1965).
17. Anderson, D. D., Quintero, C. M. & Stover, P. J. Identification of a de novo thymidylate biosynthesis pathway in mammalian mitochondria. *Proc. Natl Acad. Sci. USA* **108**, 15163–15168 (2011).
18. Kozak, M. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. *Microbiol. Rev.* **47**, 1–45 (1983).
19. Tucker, E. J. *et al.* Mutations in MTFMT underlie a human disorder of formylation causing impaired mitochondrial translation. *Cell Metab.* **14**, 428–434 (2011).
20. Saada, A. *et al.* Mutant mitochondrial thymidine kinase in mitochondrial DNA depletion myopathy. *Nat. Genet.* **29**, 342–344 (2001).
21. Calvo, S. E. & Mootha, V. K. The mitochondrial proteome and human disease. *Annu. Rev. Genomics Hum. Genet.* **11**, 25–44 (2010).
22. Agris, P. F., Vendeix, F. A. & Graham, W. D. tRNA's wobble decoding of the genome: 40 years of modification. *J. Mol. Biol.* **366**, 1–13 (2007).
23. Van Haute, L. *et al.* Deficient methylation and formylation of mt-tRNA^{Met} wobble cytosine in a patient carrying mutations in NSUN3. *Nat. Commun.* **7**, 12039 (2016).
24. Pütz, J., Dupuis, B., Sissler, M. & Florentz, C. Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures. *RNA* **13**, 1184–1190 (2007).
25. Fu, Y. *et al.* The AlkB domain of mammalian ABH8 catalyzes hydroxylation of 5-methoxycarbonylmethyluridine at the wobble position of tRNA. *Angew. Chem. Int. Ed. Engl.* **49**, 8885–8888 (2010).
26. Songe-Møller, L. *et al.* Mammalian ALKBH8 possesses tRNA methyltransferase activity required for the biogenesis of multiple wobble uridine modifications implicated in translational decoding. *Mol. Cell. Biol.* **30**, 1814–1827 (2010).
27. Yasukawa, T. *et al.* Defect in modification at the anticodon wobble nucleotide of mitochondrial tRNA^{Leu} with the MERRF encephalomyopathy pathogenic mutation. *FEBS Lett.* **467**, 175–178 (2000).
28. Yasukawa, T., Suzuki, T., Ueda, T., Ohta, S. & Watanabe, K. Modification defect at anticodon wobble nucleotide of mitochondrial tRNAs^{Leu}(UUR) with pathogenic mutations of mitochondrial myopathy, encephalopathy, lactic acidosis, and stroke-like episodes. *J. Biol. Chem.* **275**, 4251–4257 (2000).
29. Suzuki, T. & Suzuki, T. A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. *Nucleic Acids Res.* **42**, 7346–7357 (2014).
30. Ghezzi, D. *et al.* Mutations of the mitochondrial-tRNA modifier MTO1 cause hypertrophic cardiomyopathy and lactic acidosis. *Am. J. Hum. Genet.* **90**, 1079–1087 (2012).
31. Kopajtich, R. *et al.* Mutations in GTPBP3 cause a mitochondrial translation defect associated with hypertrophic cardiomyopathy, lactic acidosis, and encephalopathy. *Am. J. Hum. Genet.* **95**, 708–720 (2014).
32. Moukadir, I. *et al.* Evolutionarily conserved proteins MnmE and GidA catalyze the formation of two methyluridine derivatives at tRNA wobble positions. *Nucleic Acids Res.* **37**, 7177–7193 (2009).
33. Doherty, E. A., Batey, R. T., Masquida, B. & Doudna, J. A. A universal mode of helix packing in RNA. *Nat. Struct. Biol.* **8**, 339–343 (2001).
34. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
35. Kirino, Y., Goto, Y., Campos, Y., Arenas, J. & Suzuki, T. Specific correlation between the wobble modification deficiency in mutant tRNAs and the clinical features of a human mitochondrial disease. *Proc. Natl Acad. Sci. USA* **102**, 7127–7132 (2005).
36. Grim, J., Chládek, J. & Martínková, J. Pharmacokinetics and pharmacodynamics of methotrexate in non-neoplastic diseases. *Clin. Pharmacokinet.* **42**, 139–151 (2003).
37. Mayr, J. A. *et al.* Spectrum of combined respiratory chain defects. *J. Inher. Metab. Dis.* **38**, 629–640 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank T. Pan, W. Lu, L. Chen, L. Parsons, W. Wang and T. Srikanth, and all members of the Rabinowitz laboratory. This work was supported by funding to J.D.R. from the US National Institutes of Health (NIH) (R01CA163591 and DP1DK113643) and StandUp to Cancer (SU2C-AACR-DT-20-16). G.S.D. was supported by a postdoctoral fellowship (PF-15-190-01-TBE) from the American Cancer Society. J.A.M. was supported by the science fund of the Paracelsus Medical University Salzburg (E-12/15/076-MAY). Z.G. was supported by the NIH (DP1AI124669).

Author Contributions R.J.M., G.S.D. and J.D.R. conceived the project and designed the experiments. R.J.M. and J.D.R. wrote the manuscript. R.J.M., G.S.D. and S.H.L. performed biochemical experiments. Z.G. and W.S. were involved in study design and data interpretation. W.S. and J.A.M. contributed primary patient cell lines. All authors reviewed and edited the manuscript before submission.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to J.D.R. (josh@princeton.edu).

Reviewer Information Nature thanks D. Appling, V. Gohil and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Cell lines and growth conditions. HCT116 (CCL-247) and HEK293T/17 (CRL-11268) were purchased from ATCC. Generation of a subset of clonal CRISPR–Cas9 knockout cell lines and detailed characterization has been reported previously⁹. Additional clonal knockout cell lines (Supplementary Table 6a) were established following the protocol published previously³⁸. In brief, exon-targeting guide RNAs (Supplementary Table 6b) were designed against genes of interest and cloned into an expression vector containing the double nicking Cas9 variant (Addgene). Cells were transiently transfected using Lipofectamine 2000 (Life Technologies) (HEK293T) or Eugene HD (Promega) (HCT-116) and selected for 48 h with $2\mu\text{g ml}^{-1}$ puromycin. Single clones were isolated using serial dilution into 96-well plates.

Stable SHMT2 re-expression was achieved by transfecting HEK293T knockout cell lines with NM_005412.5 cDNA (GE Healthcare) cloned into pCMV-Tag8 vector (Agilent) and selection for three weeks with $200\mu\text{g ml}^{-1}$ hygromycin B (Sigma-Aldrich). Catalytic inactive (p.Glu98Leu/p.Tyr106Phe)^{39,40} and PLP-binding deficient (p.Lys280Gln)⁴¹ mutants were obtained following the QuickChange II protocol (Agilent). Knockout and re-expression cell lines were functionally verified by immunoblotting followed by targeted genomic sequencing (Supplementary Fig. 9) and, in the case of SHMT2 cell lines, also by tracing of [2,3,3-³H]serine labelling into dTTP. *MTOI* and *MT-TL1* patient fibroblasts and controls were provided by the Department of Paediatrics, Salzburger Landeskliniken und Paracelsus Medical University, Salzburg. Studies with primary human cell lines were approved by the local ethics-committee and informed consent was obtained from all subjects. The genotypes of *MTOI*-deficient patients (GenBank NM_012123.3) were as follows: patient-A c.[1261-5T>G];[1430G>A], (p.[?];[Arg477His]); patient-B c.[1222T>A];[1222T>A], (p.[Ile408Phe];[Ile408Phe]). Patient-B *MTOI* has been reported before⁴². Both MELAS patients carried the common *MT-TL1* m.3243A>G mutation with the heteroplasmy rate reported in this study (Extended Data Fig. 9b). All cell lines tested negative for mycoplasma and were cultured in DMEM without sodium pyruvate (Sigma-Aldrich) supplemented with 10% dialysed fetal bovine serum (dFBS, GE Healthcare) in a 5% CO₂ incubator at 37 °C. No antibiotics were used.

Glucose uptake and lactate secretion. Cells were seeded in 6-well plates 24 h before the start of an experiment. After reaching 50% confluency, plates were washed with PBS (GE Healthcare) and 3 ml of fresh media was added. Glucose uptake and lactate secretion were determined using a YSI 2900D Biochemistry Analyzer (Xylem Analytics) and normalized to cell growth as determined by micro-litre packed cell volume (PCV; as measured using packed cell volume microfuge tubes from TPP).

Proliferation. Proliferation assays were conducted in 96-well plates and relative cell number was measured using resazurin sodium salt. Approximately 5,000 cells were plated in each well with $150\mu\text{l}$ DMEM supplemented with 10% dFBS. Cell growth at each day was read as fluorescence intensity using a Synergy HT plate reader (BioTek Instruments).

Oxygen consumption. Oxygen consumption rates were measured on a XF24 extracellular flux analyser (Agilent) following the manufacturer's instructions. In brief, XF24 cell culture microplates were coated with fibronectin (Sigma-Aldrich) and cells were seeded at 5×10^4 (HEK293T) and 7×10^4 (HCT116) cells per well. After reaching 70–90% confluency, cells were equilibrated for 1 h in XF assay medium supplemented with 10 mM glucose, 1 mM sodium pyruvate, and 2 mM glutamine in a non-CO₂ incubator. Oxygen consumption rates were monitored at baseline and throughout sequential injections of oligomycin (1 μM), carbonyl cyanide-4-(trifluoromethoxy)phenylhydrazone (1 μM) and rotenone/antimycin A (0.5 μM each). Data for each well were normalized to cell number as determined by CyQUANT Cell Proliferation Assay Kit (Invitrogen) and to the on-plate wild-type control. For absolute oxygen consumption rate values, the protein concentration was determined by BCA protein assay (Thermo Fisher Scientific).

Immunoblotting. Cells were cultured to sub-confluency in 6 cm plates. After removal of media, cells were rinsed with 4 °C PBS and lysed in radio-immunoprecipitation assay (RIPA) buffer with phosphatase and protease inhibitors (Roche). Lysates were cleared by 10 min centrifugation at 16,000 g and quantified using a BCA assay (Pierce). Samples were resolved by SDS–PAGE on precast gels (Bio-Rad) and transferred to a nitrocellulose membrane using the Trans-Blot Turbo system (Bio-Rad). After overnight incubation with primary antibodies, bands were visualized with horseradish peroxidase-conjugated secondary antibodies (Cell Signaling Technologies). ChemiDoc XRS+ system was used for image acquisition. The following antibodies were used according to their manufacturer's directions: anti-SHMT1 (12612), anti-SHMT2 (12762), anti-MRPL11 (2066), anti-S6RP (2217) and anti- β -actin (5125) were from Cell Signaling Technologies; anti-MTHFD2 (ab151447), anti-NDUFS4 (ab139178), anti-SDHA (ab14715) and anti-VDAC (ab14734) were from Abcam; anti-MTHFD1L (HPA029041) was from Sigma-Aldrich.

Analysis of mitochondrial specific translation. ³⁵S-labelling of mitochondrial proteins was performed following the method described previously⁴³. In brief, cells were grown on 6 cm plates for 48 h to sub-confluency. Then media was changed to DMEM with 10% dFBS without methionine (MP Biomedicals). After a 30 min incubation, cytosolic translation was inhibited by emetine hydrochloride (0.1 mg ml⁻¹, Sigma-Aldrich) and labelling was conducted for 1 h after the addition of 500 μCi [³⁵S]methionine (EasyTag L [³⁵S]-Methionine, Perkin Elmer). Protein lysates (30 μg) were then separated on a 15% polyacrylamide gel (8.3×7.3 cm) and dried using a 443 Slab Dryer (BioRad). The dried gel was exposed to a storage phosphor screen (GE Healthcare) and imaged on a Typhoon FLA 9500 (GE Healthcare). Equal sample loading was confirmed by Coomassie brilliant blue staining (BioRad).

Metabolite concentrations and labelling patterns. Cells were grown in 6-cm dishes for at least 48 h and collected at 75% confluency. Media was replaced every 24 h and additionally 6 h before collection. Metabolism was quenched and metabolites were extracted by aspirating media and immediately adding 1 ml of 80:20 methanol:water at -80°C . Plates were kept on ice, scraped and non-soluble debris was pelleted at 18,000 g for 10 min. Samples were directly analysed by hydrophilic interaction chromatography coupled with negative-mode electrospray-ionization high resolution mass spectrometry on a quadrupole-orbitrap scanning from m/z 73 to 1,000 at 1 Hz and 140,000 resolution (Q Exactive Plus, Thermo-Fisher). Liquid chromatography separation was achieved on a XBridge BEH Amide column ($2.1\text{ mm} \times 150\text{ mm}$, $2.5\mu\text{m}$ particle size, 130 \AA pore size; Waters) using a gradient of solvent A (20 mM ammonium acetate + 20 mM ammonium hydroxide) in 95:5 water:acetonitrile, pH 9.45 and solvent B (acetonitrile). Flow rate was $150\mu\text{l min}^{-1}$. The gradient was: 0 min, 85% B; 2 min, 85% B; 3 min, 60% B; 9 min, 60% B; 9.5 min, 35% B; 12 min, 35% B; 12.5 min, 0% B; 18 min, 0% B; 18.5 min, 85% B; 23 min, 85% B. Data were processed and analysed using MAVEN software⁴⁴. All isotope tracer experiments were conducted at isotopic steady state: [2,3,3-³H]serine was traced into dTTP for a minimum of 6 h and [³⁻¹³C]serine and [¹³C]methionine were traced into tRNA for a minimum of 4 days. Isotopic tracers were purchased from Cambridge Isotope Laboratories. Isotopically labelled media was prepared from scratch and supplemented with 10% dFBS.

RNA sequencing. RNA was isolated from cell lines using RNeasy Plus kit (Qiagen) according to the manufacturers' recommendation. After the depletion of ribosomal RNA, libraries were prepared according to the TruSeq Stranded Total RNA protocol (Illumina) and sequencing was performed on a HiSeq 2500 (Illumina). Analysis was performed using the Galaxy system⁴⁵ and the R software package⁴⁶. Adaptor sequences were trimmed using Cutadapt (Galaxy version 1.6)⁴⁷ and the trimmed reads were then mapped with TopHat (Galaxy Version 0.9)⁴⁸ to the GRCh38 reference using ENSEMBL version 80 genes as known splice junctions. The read counts per gene were determined using htseq-count (Galaxy Version 0.6.1galaxy1)⁴⁹ in 'union' mode. Differential expression analysis was performed in R using DESeq2 1.12.3 package⁵⁰. SHMT2-knockout gene expression (log₂ RPKM) was graphed relative to the wild-type and the re-expressed cell lines.

Mitochondrial DNA content and integrity. To analyse mitochondrial DNA content, total DNA was extracted from 7×10^6 cells using Gentra Puregene Cell Kit (Qiagen) after freezing the cell pellet at -80°C for 1 h and overnight digestion with Proteinase K (Roche Diagnostics). qPCR (Viia 7, Applied Biosystems) was performed using primers targeting the mitochondrial ND2 locus (also known as *MT-ND2* pseudogene 1) (forward: 5'-TGTTGGTTATACCCTTCCCGTACTA-3'; reverse: 5'-CCTGCAAAGATGGTAGAGTAGATGA-3') and a nuclear *ALU* repeat sequence (forward: 5'-CTTGCACTGAGCCGAGATT-3'; reverse: 5'-GAGACGGA GTCTCGCTCTGTC-3') as published earlier⁵¹. The relative mitochondrial DNA content was determined using the $\Delta\Delta C_t$ method. Each independent sample given in the figures represents the mean of 6 technical replicates.

The mitochondrial genome was screened for deletions by long range PCR with two primer pairs spanning the whole coding region (forward_1: CCAACCAAACCCAAAGAC, reverse_1: TACTGCGACATAGGGTGCTC; forward_2: CACCAGCCTAACCAGATTTC reverse_2: ttgtaccacaaatctgttcc) and products run on a 1% agarose gel⁵². DNA from a mitochondrial DNA deletion patient was used as positive control.

For mitochondrial genomic sequence analysis, mtDNA was enriched using the multiple displacement amplification strategy (REPLI-g, Qiagen) and sequencing was performed on a HiSeq 2500 after library preparation following the Nextera library prep kit protocol (Illumina). Reads were mapped to GRCh38 using Bowtie2 (Galaxy Version 0.6)⁵³ with default settings. Coverage plots were generated using DeepTools bamCoverage (Galaxy Version 2.3.6.0)⁵⁴. The data were normalized to $1 \times$ coverage using an effective genome size of 16,569. Freebayes (Galaxy Version 0.4.1)⁵⁵ with frequency-based pooled settings was used to generate the variant data. The figures were generated in R using Gviz 1.18.0⁵⁶.

The *MT-TL1* mutation load was determined using primers specifically spanning the m.3243 position for targeted enrichment (forward: 5'-AATGATACGGCGAC CACCGAGATCTACACNNNNNGCCTTCCCCGTAATGATA-3', reverse: 5'-CAAGCAGAAGACGGCAGATCGATCGTACGGAAGGGTTGTAGT-3') followed by sequencing on a MiSeq nano flow cell using a custom sequencing primer (sequence: TATTATACCCACCCACCCCAAGAAGAGGGTTGTGAAG).⁵³ Alignment to GRCh38 was performed using Bowtie2 (Galaxy Version 0.6)⁵³ at default settings and position-specific mutation load was derived from the Integrative Genomics Viewer.

Mitochondrial ribosome profiling. Development of our ribosome profiling method was based on concepts reported previously^{57–59}. For mitochondrial ribosome profiling, cell lines were grown on 15-cm plates to 70–85% confluency. Sarcosine rescue of ribosome stalling in the SHMT2/MTFHD2 double-knockout background was assessed after growth in the presence of 1 mM sarcosine for 5 days. After removal of media, plates were rapidly rinsed with ice-cold PBS containing chloramphenicol (100 µg ml⁻¹) (Sigma-Aldrich) and cycloheximide (100 µg ml⁻¹) (Sigma-Aldrich) followed by immediate immersion into liquid nitrogen. Plates were then transferred to wet ice and 1 ml of 1.5× lysis buffer was added and the lysate was collected using a cell scraper. Lysis buffer contained the following: 1.5% Triton X-100 (Sigma-Aldrich), 0.15% NP40 (Sigma-Aldrich), 1× complete phosphatase and protease inhibitors (Roche), and 30 U ml⁻¹ DNase I (Roche) in buffer base (20 mM Tris-HCl pH 7.8 (Ambion), 100 mM KCl (Ambion), 10 mM MgCl₂ (Ambion), 100 µg ml⁻¹ chloramphenicol, 100 µg ml⁻¹ cycloheximide). 1.6–1.8 ml were recovered per plate and homogenized by passing three times through a 32G needle at 4°C. Non-soluble debris was pelleted at 5,000 g for 10 min and 1,520 µl supernatant was used for digestion with 7,500 U ml⁻¹ micrococcal nuclease (Roche) after adding 40 µl SUPERaseIN (Ambion) and 5 mM CaCl₂ (Ambion). Digestion was stopped after 1 h gentle shaking at 25°C using a final concentration of 6 mM EGTA.

Buffer base was used to make 5%–45% sucrose gradients (Gradient Master, Biocomp). After cooling to 4°C, samples were separated in an ultracentrifuge using the SW-41Ti rotor at 210,000g for 2.5 h. Live UV absorption at 254 nm was used to track the mitochondrial 5S monosome enriched fractions (Extended Data Fig. 5b). The 5S fractions were pooled and mixed with 57 µl 20% SDS per millilitre sample before performing acid phenol chloroform RNA extraction. RNA was precipitated using 300 mM sodium acetate pH 5.5 and equal volume isopropanol and run on a 15% TBE-urea gel (Invitrogen) at 210 V for 1 h for size selection. Gels were stained with Sybr Gold (Invitrogen) and RNA fragments corresponding to mitochondrial ribosome footprints (approximately 28–40 nucleotides) were cut and recovered from the gel using the crush and soak method. After sodium acetate/isopropanol precipitation, library preparation was conducted following the TruSeq Ribo Profile (Illumina) protocol.

Sequencing of ribosome protected footprints (RPFs) was performed on a HiSeq 2500 in rapid mode followed by adaptor trimming using Cutadapt (Galaxy Version 1.6)⁴⁷. Reads were mapped to the human genome reference GRCh38 using BWA (Galaxy Version 0.9)⁴⁸ with those mapping to the mitochondrial protein-coding genes included in the subsequent analysis. The Plastid package⁶⁰ and customized Phyton and R⁴⁶ scripts were used for analysing mitochondrial ribosome profiling data. Alignment was performed from 3' of reads which was reported to yield superior results after digestion with Micrococcal nuclease^{61,62}. Each read, corresponding to a mitochondrial ribosome protected fragment (mtRPF), was assigned to a nucleotide position representing the respective ribosomal A-site as determined by metagenesis analysis⁶⁰. mtRPF counts were then normalized to reads per million (RPM) mapped reads within each sample and single nucleotide positions were grouped by codon index. This transformation allows for relative quantification of bound ribosomes for each nucleotide triplet along a transcript. Stalling plots were created by plotting the mean cumulative mtRPF count along each mitochondrial open reading frame.

Codons were defined as stalling sites when the normalized counts mapped to the specific codon (mtRPF codon/mtRPF gene median) exceeded 2 s.d. from all codons in the genome. The relative abundance of actively translating ribosomes (that is, not stalled) was calculated by subtracting mtRPF counts in stalled regions from the total sum of ribosome footprints for each gene as $\Sigma \text{mtRPF}_{\text{active}} = \Sigma \text{mtRPF}_{\text{total}} - \Sigma \text{mtRPF}_{\text{stalled}}$. Then the gene specific ratio was plotted as $\Sigma \text{mtRPF}_{\text{active, SHMT2}} / \Sigma \text{mtRPF}_{\text{active, WT}}$. Stalling sites specific to the SHMT2-knockout condition were identified using the ratio of occupancy at each codon position relative to wild type. Specifically, codons were defined as SHMT2-specific stalling sites when the normalized counts in the mutant relative to wild type ($\Sigma \text{mtRPF}_{\text{codon SHMT2}} / \Sigma \text{mtRPF}_{\text{codon WT}}$) exceeded 2 s.d. (or, as indicated, 3 s.d.) from this ratio as determined for all codons in the genome, and the site also met the general stalling site criterion.

To determine the relative abundance of mitochondrial ribosomes bound to each nucleotide triplet, codon-specific occupancy ratios were calculated. For each codon (codon_{i=1–64}), the gene-specific ratio between experimentally measured

ribosome density and expected density (which is proportional to codon frequency) was calculated. Codon occupancy (CO_{i=1–64}) for each codon is the mean of the ratios from all 13 genes. The relative codon occupancy (CO_{i=1–64, SHMT2}) / (CO_{i=1–64, WT}) was plotted with error bars representing s.d. across replicates after error propagation. To investigate the ribosome distribution relative to the major stalled codons (AAG and UUG), ribosome densities flanking the codons of interest within 25 amino acids were selected. Each selected fragment was first normalized to its total count so every codon of interest from the genome is weighted equally. The mean value from each position was plotted.

Investigating protein secondary structure effects on stalling at AAG and UUG.

On average stalling was most pronounced at AAG and UUG codons, but not all codons of the same sequence were equally affected. We therefore investigated whether the positioning relative to protein secondary structures (transmembrane helices) influences the extent of stalling. As micrococcal nuclease treatment induces imprecision in A-site mapping owing to sequence-biased digestion^{63,64}, individual positions identified as SHMT2-specific stalling sites were first grouped to the adjacent codons decoded by 5-taurinomethyluridine-modified tRNAs. Then amino acid residues corresponding to the AAG and UUG codons were mapped to *Bos taurus* crystal structures of mitochondrial respiratory chain complex proteins using iCn3D⁶⁵. Structure data was retrieved using the following Protein Data Bank (PDB) accession codes: ATP6: 5ARA_W; MT-CO2: 2Y69_B; MT-CYB: 1QCR_C; and MT-ND6: 5LDW_J (underscored suffixes denote the respective subunits of the supercomplexes). In addition, a hidden Markov model based algorithm for transmembrane helices (TMHMM 2.0⁶⁶) was used to predict α-helical transmembrane domains in the *Homo sapiens* sequences. This method assigns each codon a probability for transmembrane helix localization which was then used for genome wide assessment of AAG and UUG localization relative to transmembrane helices. AAG and UUG codons were defined to be at a transition between a transmembrane helix and a non-membrane region if, within the five flanking codons, probabilities >0.5 and <0.5 for being in a transmembrane helix are found. AAG and UUG codons were defined as stalling sites based on the 3 s.d. cut-off as per Extended Data Fig. 8b. In total, 4 out of 5 stalling AAG and UUG codons, and 7 out of 23 non-stalling AAG and UUG codons were at a membrane transition ($P=0.04$ by chi-square test).

Evaluation of mRNA secondary structure effects on ribosome stalling.

To study a potential effect of mRNA secondary structure⁶⁷ (that is, base pairing) on ribosome stalling, we used the previously published dimethyl sulfate sequencing datasets on human K562 cell lines³⁴ to identify structured regions in mitochondrial transcripts. Following the methods described in the manuscript for nuclear transcripts, identification of sites with secondary structure was performed on mitochondrial transcript data³⁴. In brief, FASTQ files (accession numbers GSM1297495 and GSM1297493) were retrieved from sequence read archive and mapped to GRCh38 with BWA (Galaxy Version 0.9)⁴⁸. Reads were assigned to the nucleotide at the 5' end with no offset using Plastid⁶⁰. R values (cut-off 0.75) and Gini differences (cut-off 0.1) between the *in vivo* and denatured dataset were calculated for the complete mitochondrial transcriptome for a window size of 50 adenosine/cytosine nucleotides and a step size of 10. This provided a list of structured mitochondrial transcript regions, with most mRNA regions unstructured. The list of structured regions was compared to the SHMT2-specific stalling sites for potential co-localization. No stalling site mapped to a structured region.

Mitochondrial enzyme activities. Activities of individual OXPHOS complexes I–IV, ATP synthase and citrate synthase (which is nuclear encoded and was used as a marker of mitochondrial mass) were spectrophotometrically measured as previously described (Uvicon 922, Kontron)^{68–70}. Measurements were performed with 2 µl of mitochondria isolated by differential centrifugation⁷¹ (except for complex I, in which 10 µl was used).

Citrate synthase (EC 2.3.3.1) activity was determined following extinction dynamics at 412 nm, indicating the cleavage of Elman's reagent (0.2 mM) after addition of oxaloacetate (0.5 mM) to the buffered reaction solution containing acetyl-CoA (0.15 mM). Rotenone-sensitive complex I (NADH:decylubiquinone oxidoreductase, EC 1.6.5.3) activity was measured by adding NADH (0.2 mM) and monitoring at 340 nm for the reduction of decyl-ubiquinone (50 µM). Complex II (succinate:ubiquinone-oxidoreductase, EC 1.3.5.1) was measured at 600 nm by monitoring the reduction of 2,6-dichlorophenol-indophenol (80 µM) after addition of succinate (10 mM). The reaction mixture to determine complex III activity (coenzyme Q:cytochrome c-oxidoreductase, EC 1.10.2.2) contained cytochrome c (100 µM) and decyl-ubiquinol (200 µM) and was measured at 550 nm. After inhibition by addition of antimycin A (1 µM), the insensitive activity was subtracted to calculate specific complex III activity. The enzyme activity of complex IV (ferrocytochrome c: oxygen oxidoreductase, EC 1.9.3.1) was read as the oxidation rate of reduced cytochrome C (60 µM) at 550 nm. Complex V (F₁F₀ ATP synthase, EC 3.6.3.14) was indirectly measured as oligomycin-sensitive ATPase activity in a reaction mixture containing 0.5 mM ATP. Formed ADP was coupled

to a pyruvate kinase reaction, using phosphoenolpyruvate (2 mM) to generate ATP and pyruvate. The latter is then used by lactate dehydrogenase in the oxidation of NADH (0.2 mM) that served as readout (340 nm). Reagents were obtained from Sigma-Aldrich.

Mitochondrial tRNA modifications. HCT116 or HEK293T cell lines and sub-clones were grown to 70–85% confluency and collected for mitochondrial extraction (1×10^8 – 2×10^8)⁷¹. Mitochondrial tRNAs were extracted using the MirVana miRNA Isolation kit (Ambion) for isolation of small RNAs followed by 10% TBE-urea gel purification and extraction as described above (tRNA fraction 65–85 base pairs). Quantitative analysis of 5-taurinomethyluridine monophosphate (p- τ^m s²U), 5-taurinomethyl-2-thiouridine monophosphate (p- τ^m s²U), 2-thiouridine monophosphate (p-s²U) and 5-formylcytidine monophosphate (p-f²C) was performed using high-resolution mass spectrometry following an adapted previously published protocol^{25,72}. In brief, 100 ng tRNA were digested at 37°C for 2 h by nuclease P1 (2 U) in 30 μ l of 100 mM ammonium acetate and quenched by adding 60 μ l 50/50 methanol/acetonitrile followed by centrifugation at 10,000 g for 10 min. 5 μ l of sample was injected for LC-MS analysis. Nucleoside monophosphates were analysed on a quadrupole-orbitrap mass spectrometer (Q Exactive plus, Thermo Fisher Scientific) operating in negative-ion mode coupled to hydrophilic interaction chromatography via electrospray ionization and used to scan in SIM mode from m/z 459 to 482 (τ^m s²U and τ^m s²U) or m/z 338 to 355 (s²U and f²C) at 1 Hz and 140,000 resolution. Consistent loading was ensured by measuring f²C levels (which are not altered by SHMT2 knockout) in the same sample as the τ^m s²U and τ^m s²U. The mass spectrometry standard for 5-taurinomethyluridine was synthesized as described previously⁷³ and was used to confirm peak identity in digested tRNA samples after an additional treatment with 1 U of alkaline phosphatase (Roche) in 100 mM ammonium carbonate at 37°C for 2 h⁷².

Folate depletion and methotrexate treatment. Cells were grown for five passages in folic acid-deficient DMEM with 10% dFBS (US Bio), either with or without hypoxanthine and thymidine supplementation (Gibco; final concentrations of 100 μ M sodium hypoxanthine and 16 μ M thymidine). To evaluate the effect of targeting 1C-metabolism on mitochondrial function, methotrexate (Sigma-Aldrich) was used at 25, 50 and 75 nM concentration. Chloramphenicol (100 μ g ml⁻¹ = 310 μ M) and ethidium bromide (100 ng ml⁻¹ = 250 nM) served as positive controls for the inhibition of mitochondrial translation and mtDNA depletion respectively. At each time point, protein and DNA samples were collected for immunoblot and mtDNA content analysis. Basal respiration was assessed using the Seahorse XF analyser with measurements conducted as described above.

N-terminal protein formylation. N-formylmethionine modification on COX1 was assayed by mass spectrometry following the protocol described previously¹⁹. In brief, mitochondria were isolated from cells by differential centrifugation⁷¹ and complex IV was immunoprecipitated (ab109801, Abcam) before separation on a 4–20% polyacrylamide gel. A band running at the same molecular mass as a band reactive with an anti-COX1 antibody (ab14705, Abcam) was excised and lysed using 1.5 μ g LysC (Wako) as described⁷⁴. Samples were dried in a speedvac and re-suspended with 15 μ l of 0.1% formic acid pH 3. Per run, 5 μ l was injected using an Easy-nLC 1000 UPLC system. Samples were loaded directly onto a 45 cm \times 75 μ m nano-capillary column packed with 1.9 μ m C18-AQ (Dr. Maisch) mated to a metal emitter (Thermo Scientific) in-line with a Thermo Orbitrap Elite or Thermo Orbitrap Lumos. The mass spectrometer was operated in data-dependent mode with the 120,000 resolution MS1 scan (400–1,800 m/z) in the Orbitrap followed by up to 20 MS/MS scans in the ion trap. Raw files were searched using MS Amanda⁷⁵, Sequest HT⁷⁶ and Byonic⁷⁷ algorithms and validated using the Percolator algorithm⁷⁸ within the Proteome Discoverer 2.1 suite (Thermo Scientific). 10 p.p.m. MS1 and 0.6 Da MS2 mass tolerances were specified. Carbamidomethylation of cysteine was used as fixed modification and oxidation of methionine as dynamic modification. In addition, acetylation, formylation and loss of methionine were specified as potential modifications at the N terminus of proteins. The resulting msf file was used to construct a spectral library (percolator peptide $q > 0.95$) and extract MS1 ion chromatographs in Skyline^{79,80}. The fraction of modified N-terminal peptides of COX1 was calculated as the area of the formylated peptide divided by the sum of the areas of all the n-terminal peptides in that sample.

Statistics and reproducibility. Significance was determined by two-tailed Student's *t*-test comparing the indicated condition to the corresponding wild type or control. Asterisks denote $P < 0.01$. Exact *P* values for individual comparisons are given in Supplementary Table 7. Small filled circles are individual data points. All results have been independently replicated at least twice. This includes figure panels where representative data are shown. For quantitative measurements, *n* is provided in the figure legends. For western blots and DNA gels, we show representative data of multiple independent replicates: Fig. 1c (*n* = 2), Fig. 2b (*n* = 3), Fig. 4c (*n* = 3); Extended Data Fig. 1a (*n* = 2), Extended Data Fig. 1h (*n* = 2), Extended Data Fig. 1i (*n* = 2), Extended Data Fig. 2b (*n* = 2), Extended Data Fig. 2c (*n* = 2),

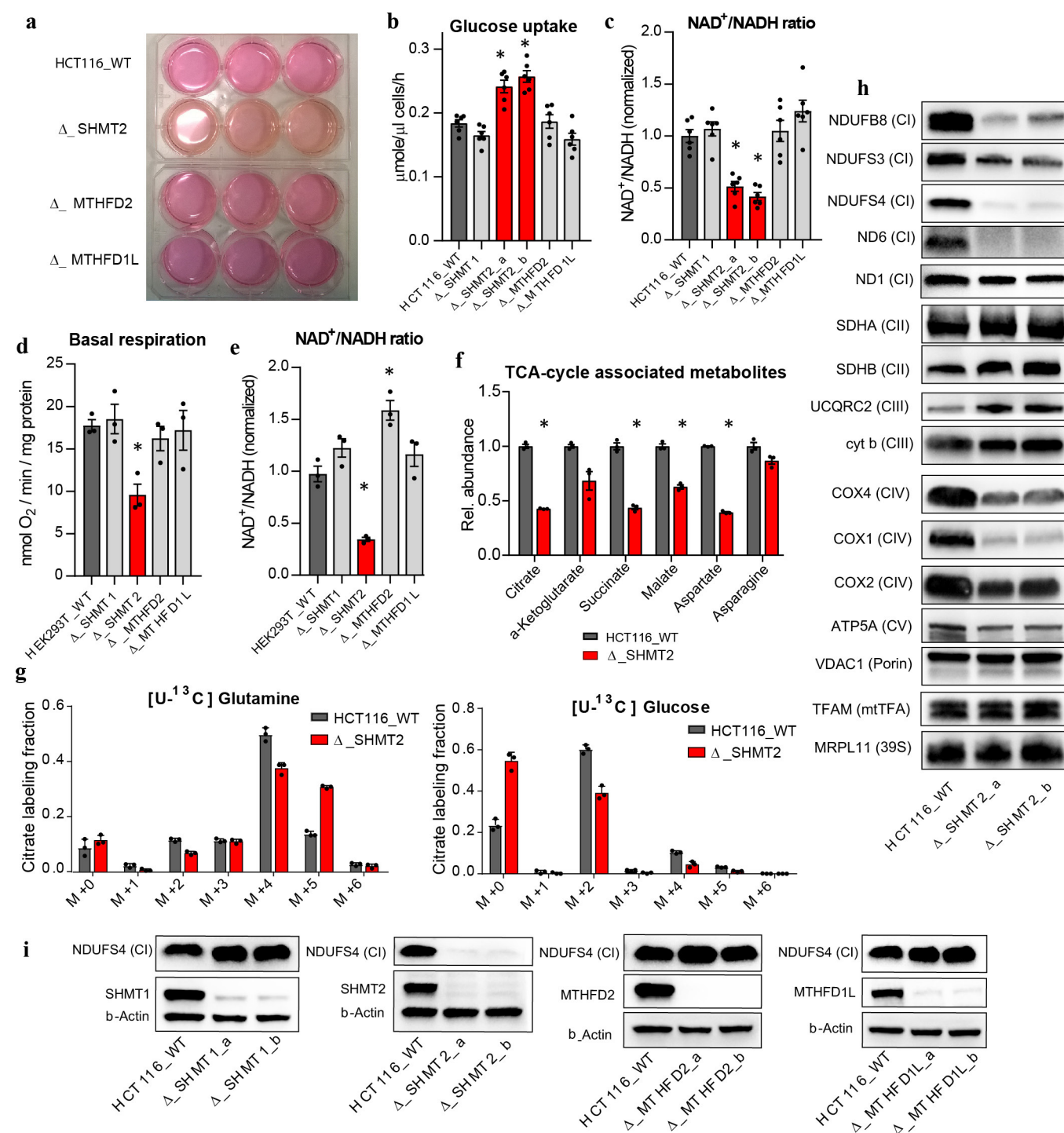
Extended Data Fig. 3a (*n* = 2), Extended Data Fig. 4d (*n* = 2), Extended Data Fig. 5a (*n* = 2), Extended Data Fig. 5b (*n* = 3), Extended Data Fig. 7f (*n* = 2), Extended Data Fig. 10a (*n* = 3) and Extended Data Fig. 10c (*n* = 2). No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Code availability. All code used to generate the data in this manuscript is publicly available from GitHub (<https://github.com/R-J-Morscher>).

Data availability. Raw sequencing data are available from Sequence Read Archive (SRA) under BioProject accession number PRJNA419990. Source data for Figs 1c, 3c, 4c, Extended Data Figs 3d, 6e are provided with the paper. Uncropped versions of blots are provided in Supplementary Figs 1–8; Sanger sequencing traces are provided in Supplementary Fig. 9. Other data that support the findings of this study are available from the corresponding author upon request.

38. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protocols* **8**, 2281–2308 (2013).
39. Szebenyi, D. M., Musayev, F. N., di Salvo, M. L., Safo, M. K. & Schirch, V. Serine hydroxymethyltransferase: role of glu75 and evidence that serine is cleaved by a retroaldol mechanism. *Biochemistry* **43**, 6865–6876 (2004).
40. Contestabile, R. *et al.* Role of tyrosine 65 in the mechanism of serine hydroxymethyltransferase. *Biochemistry* **39**, 7492–7500 (2000).
41. Iurescia, S., Condò, I., Angelaccio, S., Delle Fratte, S. & Bossa, F. Site-directed mutagenesis techniques in the study of *Escherichia coli* serine hydroxymethyltransferase. *Protein Expr. Purif.* **7**, 323–328 (1996).
42. Tischner, C. *et al.* MTO1 mediates tissue specificity of OXPHOS defects via tRNA modification and translation optimization, which can be bypassed by dietary intervention. *Hum. Mol. Genet.* **24**, 2247–2266 (2015).
43. Sasarman, F. & Shoubridge, E. A. Radioactive labeling of mitochondrial translation products in cultured cells. *Methods Mol. Biol.* **837**, 207–217 (2012).
44. Clasquin, M. F., Melamud, E. & Rabinowitz, J. D. LC-MS data processing with MAVEN: a metabolomic analysis and visualization engine. *Curr. Protoc. Bioinformatics* **Chapter 14**, Unit14.11 (2012).
45. Afgan, E. *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.* **44** (W1), W3–W10 (2016).
46. R Development Core Team. *R: A Language and Environment for Statistical Computing*; <http://www.R-project.org/> (Vienna, Austria, 2016).
47. Marcel, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
49. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
51. Bao, X. R. *et al.* Mitochondrial dysfunction remodels one-carbon metabolism in human cells. *eLife* **5**, e10575 (2016).
52. Mayr, J. A. *et al.* Mitochondrial phosphate-carrier deficiency: a novel disorder of oxidative phosphorylation. *Am. J. Hum. Genet.* **80**, 478–484 (2007).
53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
54. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44** (W1), W160–W165 (2016).
55. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. Preprint at <https://arxiv.org/abs/1207.3907> (2012).
56. Hahne, F. & Ivanek, R. Visualizing genomic data using Gviz and Bioconductor. *Methods Mol. Biol.* **1418**, 335–351 (2016).
57. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protocols* **7**, 1534–1550 (2012).
58. Rooijers, K., Loayza-Puch, F., Nijtmans, L. G. & Agami, R. Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat. Commun.* **4**, 2886 (2013).
59. Couvillion, M. T., Soto, I. C., Shipkova, G. & Churchman, L. S. Synchronized mitochondrial and cytosolic translation programs. *Nature* **533**, 499–503 (2016).
60. Dunn, J. G. *plastid: a positional library for sequencing analysis*; <http://plastid.readthedocs.io> (2016).
61. Nakahigashi, K. *et al.* Effect of codon adaptation on codon-level and gene-level translation efficiency in vivo. *BMC Genomics* **15**, 1115 (2014).
62. Balakrishnan, R., Oman, K., Shoji, S., Bundschuh, R. & Fredrick, K. The conserved GTPase LepA contributes mainly to translation initiation in *Escherichia coli*. *Nucleic Acids Res.* **42**, 13370–13383 (2014).
63. Oh, E. *et al.* Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* **147**, 1295–1308 (2011).
64. Dunn, J. G., Foo, C. K., Belletier, N. G., Gavis, E. R. & Weissman, J. S. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila melanogaster*. *eLife* **2**, e01179 (2013).
65. Wang, Y., Geer, L. Y., Chappay, C., Kans, J. A. & Bryant, S. H. Cn3D: sequence and structure views for Entrez. *Trends Biochem. Sci.* **25**, 300–302 (2000).
66. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

67. Del Campo, C., Bartholomäus, A., Fedyunin, I. & Ignatova, Z. Secondary structure across the bacterial transcriptome reveals versatile roles in mRNA regulation and function. *PLoS Genet.* **11**, e1005613 (2015).
68. Danson, M. J. & Hough, D. W. Citrate synthase from hyperthermophilic Archaea. *Methods Enzymol.* **331**, 3–12 (2001).
69. Feichtinger, R. G. *et al.* Low aerobic mitochondrial energy metabolism in poorly- or undifferentiated neuroblastoma. *BMC Cancer* **10**, 149 (2010).
70. Rustin, P. *et al.* Biochemical and molecular investigations in respiratory chain deficiencies. *Clin. Chim. Acta* **228**, 35–51 (1994).
71. Clayton, D. A. & Shadel, G. S. Isolation of mitochondria from tissue culture cells. *Cold Spring Harb. Protoc.* <http://doi.org/10.1101/pdb.prot080002> (2014).
72. Zheng, G. *et al.* ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell* **49**, 18–29 (2013).
73. Ogata, T. *et al.* Chemical synthesis and properties of 5-taurinomethyluridine and 5-taurinomethyl-2-thiouridine. *J. Org. Chem.* **74**, 2585–2588 (2009).
74. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protocols* **1**, 2856–2860 (2006).
75. Dorfer, V. *et al.* MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra. *J. Proteome Res.* **13**, 3679–3684 (2014).
76. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–989 (1994).
77. Bern, M., Kil, Y. J. & Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* **Chapter 13**, Unit13.20 (2012).
78. Spivak, M., Weston, J., Bottou, L., Käll, L. & Noble, W. S. Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets. *J. Proteome Res.* **8**, 3737–3745 (2009).
79. MacLean, B. *et al.* Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).
80. Schilling, B. *et al.* Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol. Cell. Proteomics* **11**, 202–214 (2012).
81. Renwick, S. B., Snell, K. & Baumann, U. The crystal structure of human cytosolic serine hydroxymethyltransferase: a target for cancer chemotherapy. *Structure* **6**, 1105–1116 (1998).

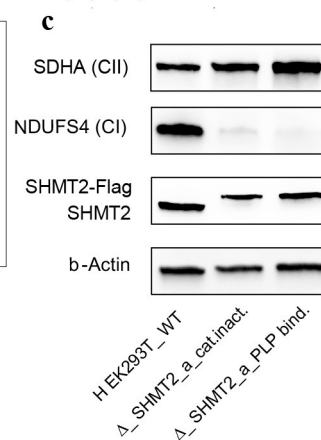
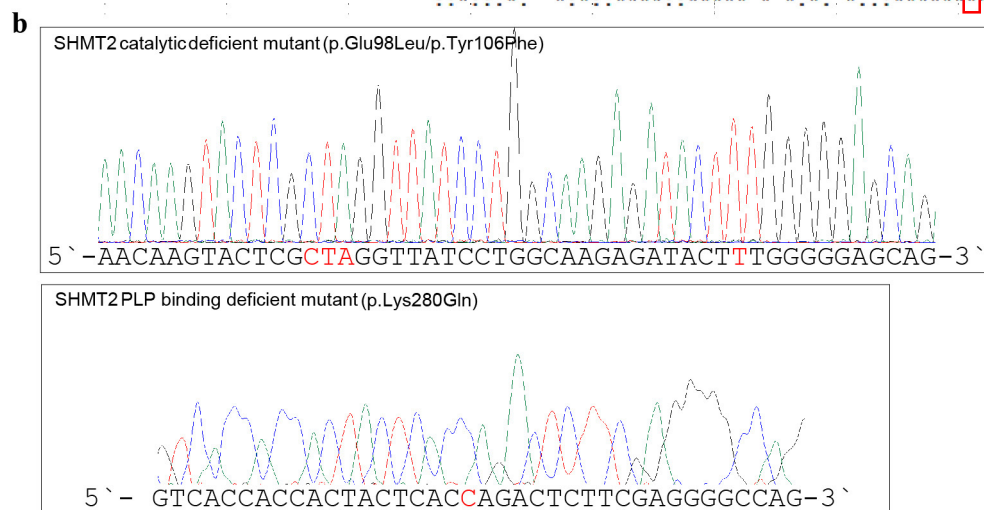
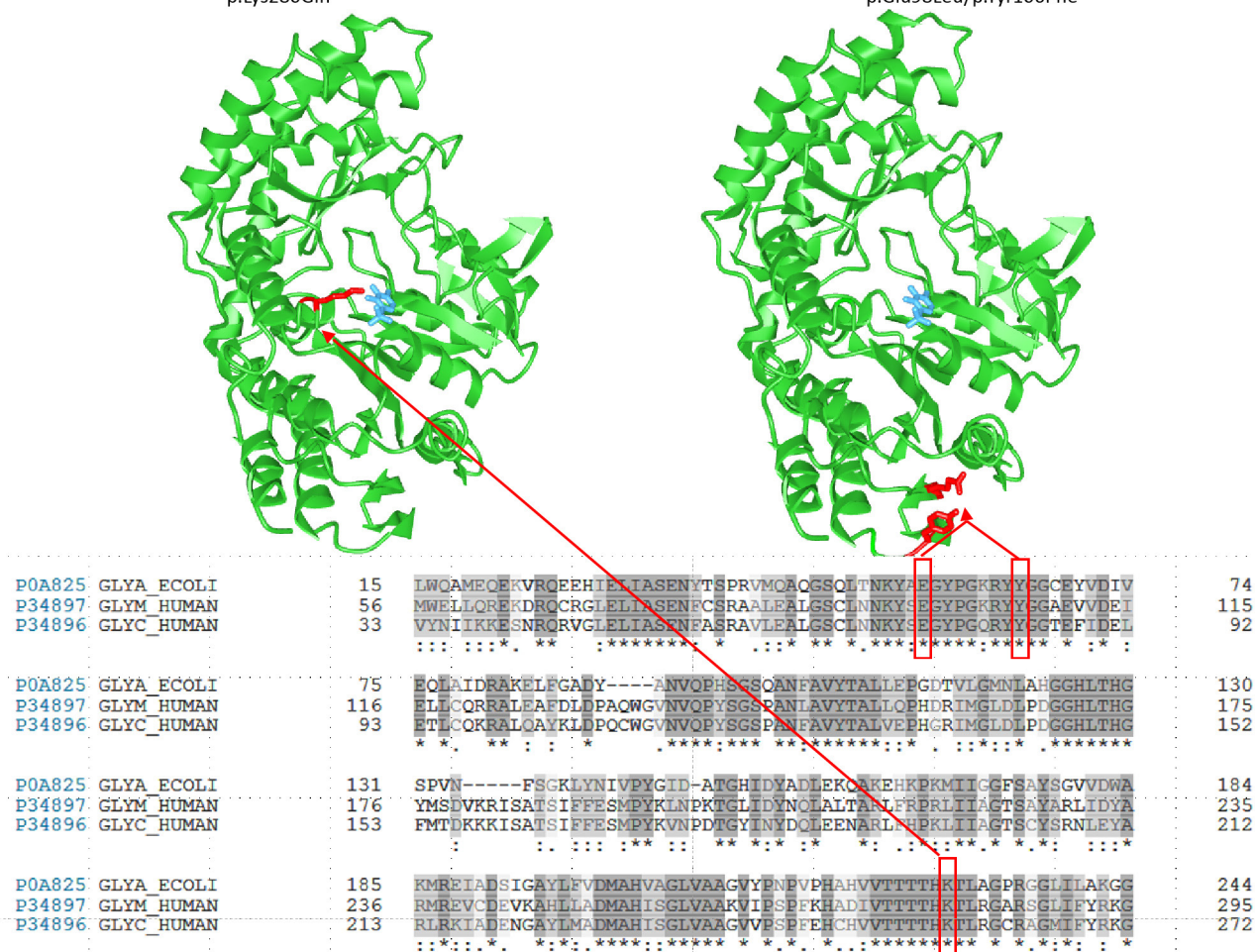


Extended Data Figure 1 | SHMT2 deletion-induced respiratory chain dysfunction in different cellular backgrounds and clones. a, Change in media colour after 48 h cell growth. **b, c,** Lactate secretion (**b**) and normalized NAD⁺/NADH ratio (**c**) of HCT116 knockout cell lines ($n = 6$). **d, e,** Basal respiration as measured by Seahorse XF analyser ($n = 3$) (**d**) and normalized NAD⁺/NADH ratio ($n = 3$) (**e**) of HEK293T folate 1C gene CRISPR–Cas9 knockout cell lines. **f,** Normalized levels of TCA cycle and associated metabolites ($n = 3$). **g,** Steady-state labelling fraction into citrate

from [U-¹³C] substrates glutamine (left) and glucose (right) ($n = 3$). **h,** Immunoblot of extracted mitochondria for subunits of respiratory chain complexes I–V (CI–CV) and markers of mitochondrial mass. **i,** Mitochondrial complex I levels (NDUFS4) in independent HCT116 folate 1C gene knockout clones. Data are mean \pm s.e.m. n indicates the number of biological replicates. $*P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values).

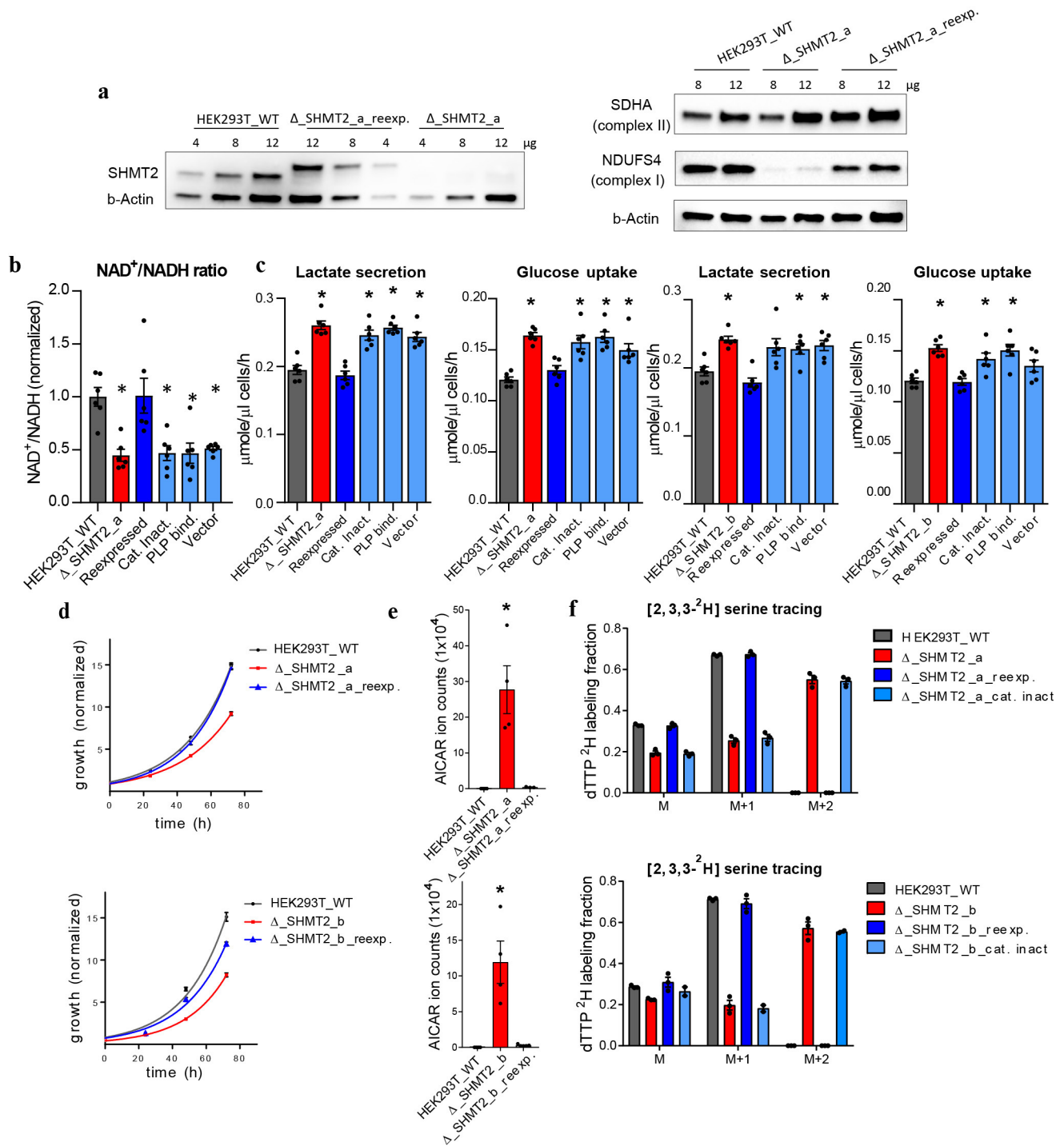
a SHMT2 pyridoxal phosphate binding deficient mutant
p.Lys280Gln

SHMT2 catalytic deficient double mutant
p.Glu98Leu/p.Tyr106Phe



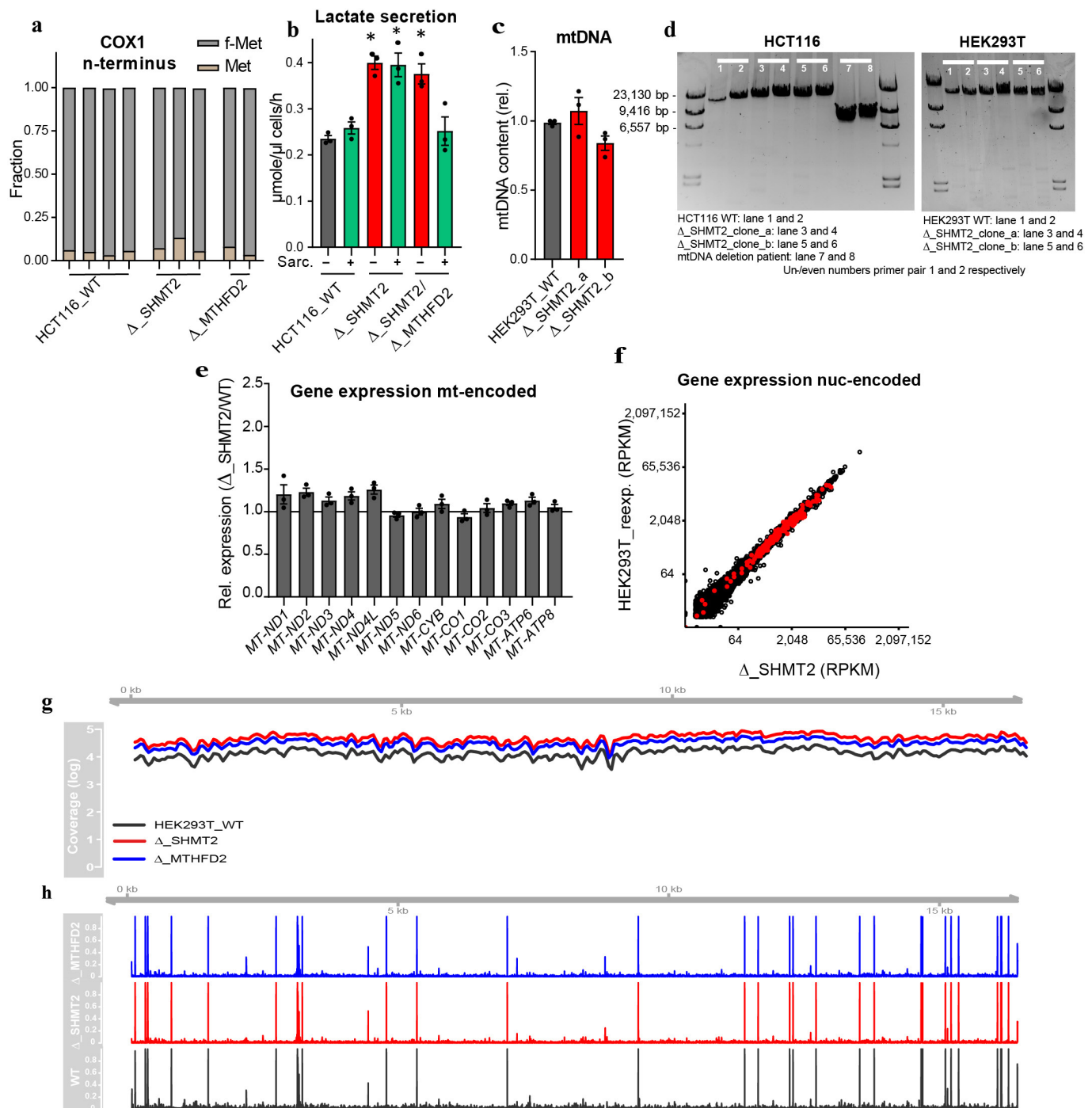
Extended Data Figure 2 | Catalytically deficient SHMT2 constructs. **a**, Mapping of mutated amino acid residues on human SHMT1 (PDB code 1BJ4⁸¹) using iCn3D and alignment of *E. coli* serine hydroxymethyltransferase (GLYA), *H. sapiens* mitochondrial serine hydroxymethyltransferase 2 (GLYM) and cytosolic serine

hydroxymethyltransferase 1 (GLYC). Positions for GLYM are given with reference to GenBank NM_005412.5. **b**, Sanger sequencing traces of mutant constructs. **c**, Immunoblot for mitochondrial complex I levels (NDUFS4) in cell lines re-expressing catalytically deficient forms of SHMT2.



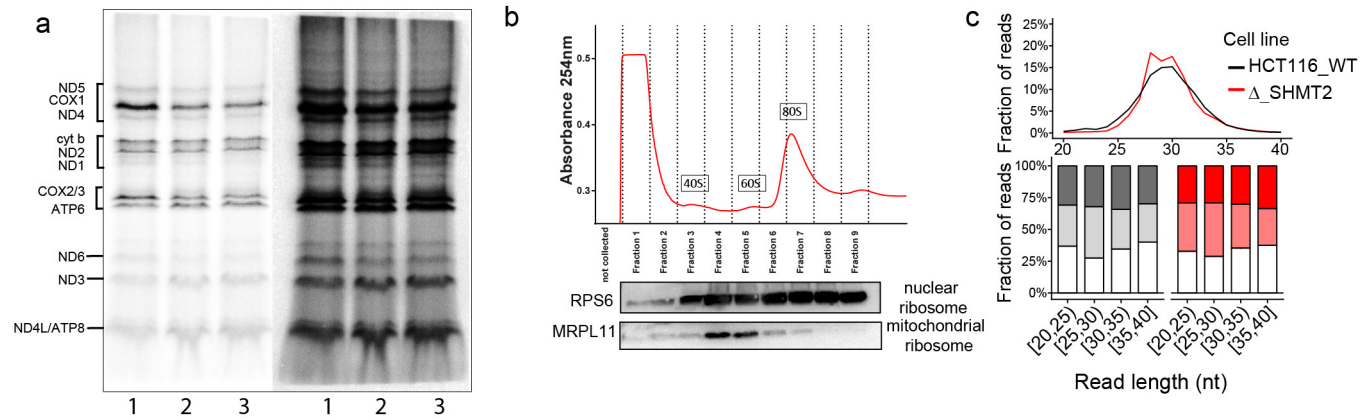
Extended Data Figure 3 | Restoring SHMT2 catalytic activity normalizes 1C flux, respiratory chain expression, glycolytic activity, and cell growth. **a**, Immunoblot of re-expression of catalytically active SHMT2 (left) and the effects of its re-expression on mitochondrial complex I and II levels (right). **b–f**, Effect of re-expression of catalytically active and inactive forms of SHMT2 in two different Δ SHMT2 clones in the HEK293T background. **b**, Normalized NAD⁺/NADH ratio ($n=6$). **c**, Lactate secretion and glucose uptake ($n=6$). **d**, Cell proliferation

($n=6$). **e**, Purine biosynthesis intermediate 5-aminoimidazole-4-carboxamide ribonucleotide (AICAR) levels ($n=4$) as an indicator of cytosolic folate 1C status. **f**, [2,3,3-²H]serine tracing to differentiate cytosolic from mitochondrial folate 1C unit production for incorporation into deoxythymidine triphosphate ($n=3$). Data are mean \pm s.e.m. n indicates the number of biological replicates. * $P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values).



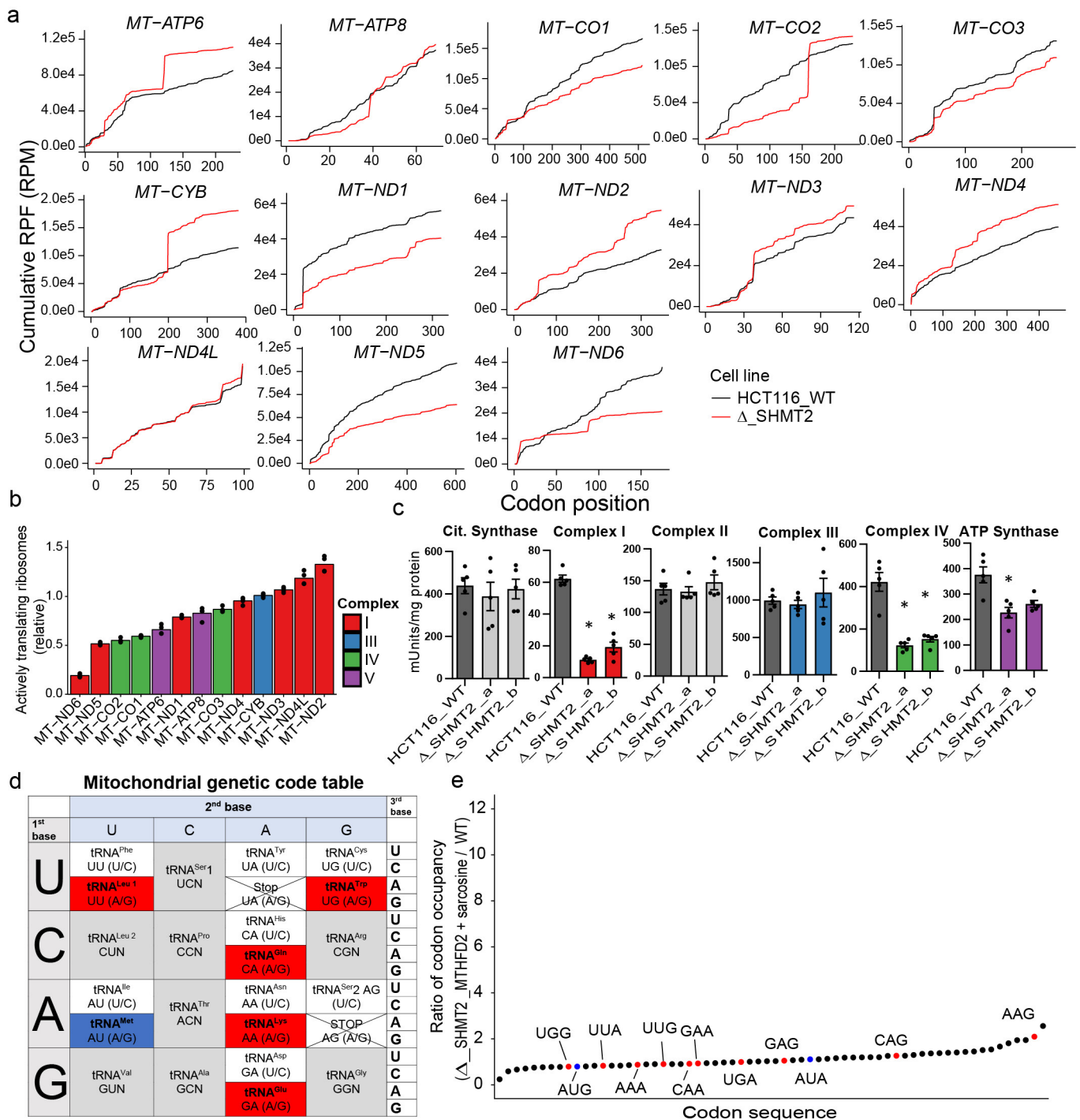
Extended Data Figure 4 | Oxidative phosphorylation defect is caused by a post-transcriptional mechanism independent of methionine formylation. **a**, Fraction of initiating amino acid (formylmethionine versus methionine) of mitochondrial-expressed COX1 peptide determined by high-resolution LC-MS (wild type $n=4$, Δ SHMT2 $n=3$, Δ MTHFD2 $n=2$). **b**, Lactate secretion ($n=3$) upon sarcosine supplementation (1 mM). **c**, Relative mtDNA levels in HEK293T cells ($n=3$). **d**, Agarose gel of mtDNA long-range PCR products of HCT116 and HEK293T knockout cell lines. **e**, Relative mRNA levels of mtDNA-encoded respiratory chain subunits in the HEK293T background ($n=3$). **f**, Gene expression levels in SHMT2-knockout cell lines compared to SHMT2 wild-type re-expressed

lines by total RNA sequencing. Each dot represents mean gene expression as derived from two biological replicates of two independent knockout clones and matched re-expressed lines ($n=4$). Genes linked to human OXPHOS function³⁷ are highlighted in red. Significantly differentially expressed genes are listed in Supplementary Table 2. **g**, Position-dependent next-generation sequencing coverage of mtDNA in HEK293T wild-type, SHMT2-knockout and MTHFD2-knockout cell lines supports the absence of deletions due to SHMT2 loss. **h**, Corresponding variant position and frequency. Variant list is provided in Supplementary Table 1. Data are mean \pm s.e.m. n indicates the number of biological replicates. $*P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values).



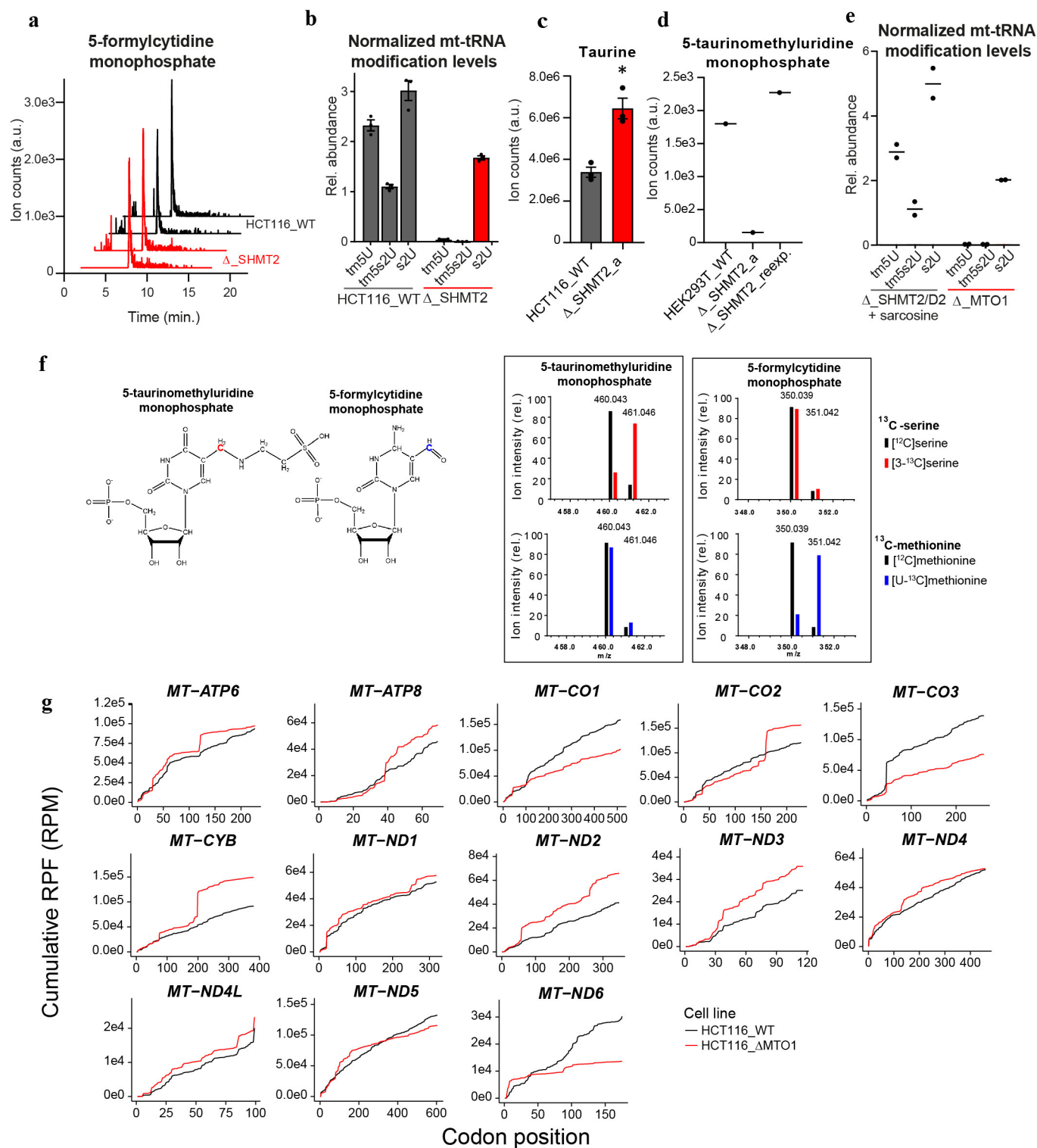
Extended Data Figure 5 | Impairment of mitochondrial translation due to loss of SHMT2. **a**, SDS-PAGE of $[^{35}\text{S}]$ methionine-labelled mitochondrially translated proteins in wild-type (lane 1) and two SHMT2-knockout (lane 2 and 3) HEK293T cell lines. Decreased synthesis of COX1 and COX2/3 are evident upon short exposure and reduced synthesis of ND5 and ND6 is more easily visualized upon longer exposure. **b**, Absorbance at 254 nm upon sucrose gradient fractionation of cell lysates digested by micrococcal nuclease (Fig. 3a). Fractions corresponding to 4

and 5 were collected for mitochondrial ribosome enrichment as shown on the matched immunoblot for mitochondrial ribosome subunit MRPL11. **c**, Read length distribution (top) and read length-dependent sub-codon read phasing (bottom) across the 13 mitochondrial protein-coding transcripts. Data in **c** are based on the mitochondrial ribosome profiling experiment in Fig. 3, and represent the mean of two technical replicates of two independent samples.



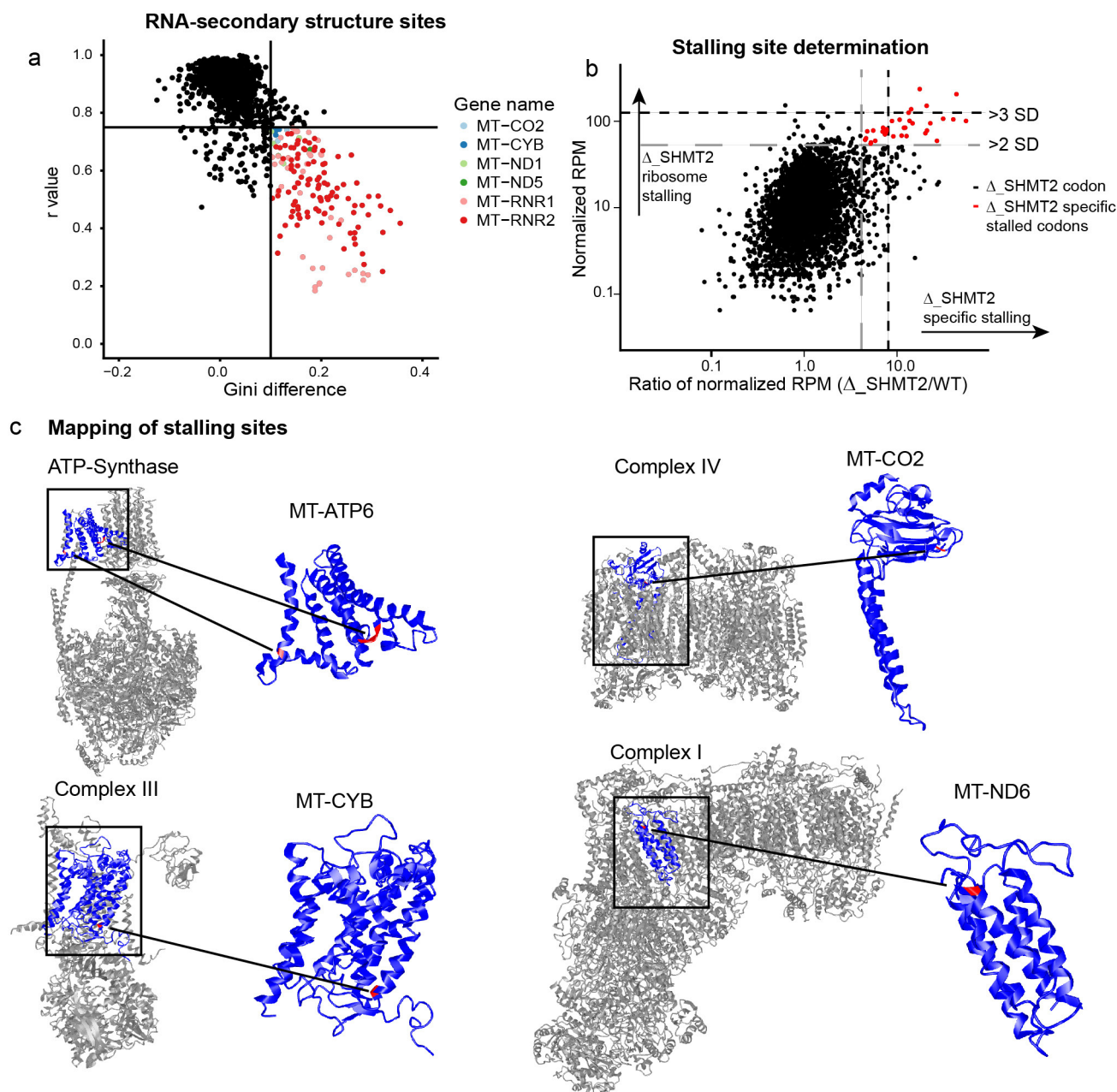
Extended Data Figure 6 | Mitochondrial ribosome stalling at guanosine-ending split codon box nucleotide triplets suggests deficient 5-taurinomethyluridine modification. **a**, Expanded version of Fig. 3b, showing the mean cumulative ribosome protected fragments of all mitochondrial protein-coding genes. **b**, Mean relative density of actively translating (that is, not stalled) ribosomes for mitochondrial transcripts. Data in **a** and **b** represent two technical replicates of two independent samples. **c**, Enzymatic activities of citrate synthase and individual mitochondrial respiratory chain complexes from mitochondrial extracts ($n = 5$). Data are mean \pm s.e.m. * $P < 0.01$, two-tailed Student's t -test

(see Supplementary Table 7 for exact P values). **d**, Mitochondrial genetic code table with split codon boxes depending on taurinomethylated tRNAs for translation highlighted in red. Codons decoded by anticodon formylcytidine-containing tRNA^{Met} are highlighted in blue. **e**, Mean codon-specific mitochondrial ribosome occupancy of HCT116 SHMT2/MTHFD2 double-knockout cell lines supplemented with sarcosine (1 mM). Codons highlighted in red are decoded by tRNAs carrying a 5-taurinomethyluridine modification. The supplementation with sarcosine prevents the stalling normally observed with SHMT2 deletion ($n = 2$).



Extended Data Figure 7 | tRNA modification status in Δ SHMT2 and effects of 5-taurinomethyluridine modification loss caused by human disease gene *MTO1*. **a**, Total ion chromatogram of 5-formylcytidine monophosphate in digested mitochondrial tRNAs upon loss of SHMT2. The same samples were analysed for 5-taurinomethyluridine monophosphate (p- τ m⁵U) in Fig. 4b. The combined data demonstrate that SHMT2 deletion causes loss of τ m⁵U but not 5-formylcytidine. **b**, Levels of τ m⁵U, 5-taurinomethyl-2-thiouridine monophosphate (p- τ m⁵s²U) and 2-thiouridine monophosphate (p-s²U) in wild-type HCT116 and SHMT2 deletion lines normalized to 5-formylcytidine monophosphate (p-f⁵C) ($n = 3$). **c**, Taurine levels in HCT116 wild-type and SHMT2-knockout cells ($n = 3$). **d**, τ m⁵U levels in digested mitochondrial tRNAs upon

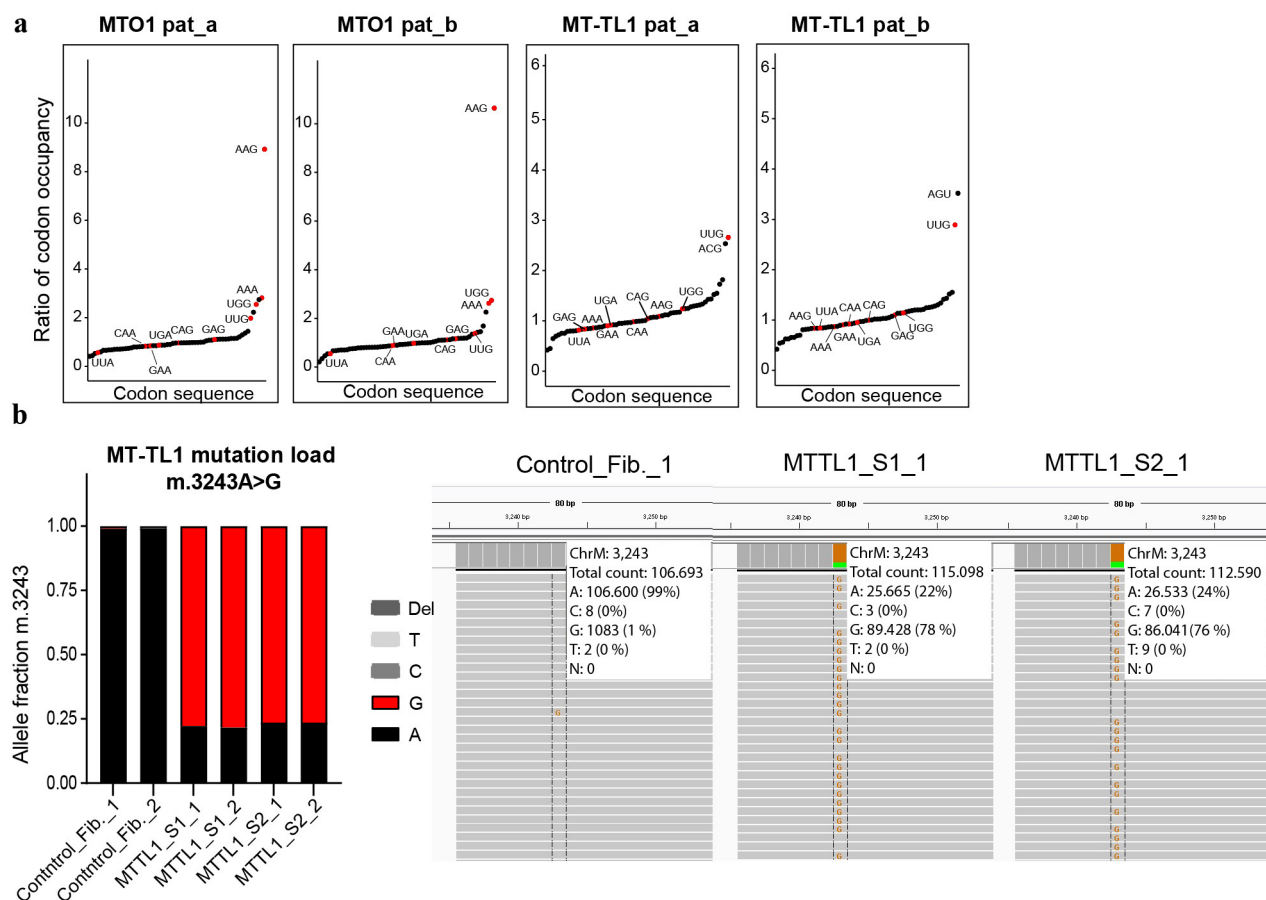
re-expression of SHMT2 ($n = 1$). **e**, τ m⁵U, τ m⁵s²U and s²U levels normalized to f⁵C in HCT116 SHMT2/MTHFD2 knockout lines after sarcosine supplementation and HCT116 upon loss of MTO1 ($n = 2$). For all panels, data are mean \pm s.e.m. or individual data points only. **f**, Labelling pattern of 5-taurinomethyluridine and 5-formylcytidine monophosphate extracted from mitochondrial tRNAs after growth in media containing either [3-¹³C]serine or [U-¹³C]methionine. **g**, Mean cumulative count of ribosome protected fragments (RPF) mapping to mitochondrial protein coding transcripts upon ribosome profiling in HCT116 MTO1-knockout cell lines. Data were normalized to RPM ($n = 2$); n indicates the number of biological replicates. * $P < 0.01$, two-tailed Student's *t*-test (see Supplementary Table 7 for exact *P* values).



Extended Data Figure 8 | Investigation of mRNA and protein secondary structure effects on mitochondrial ribosome stalling sites.

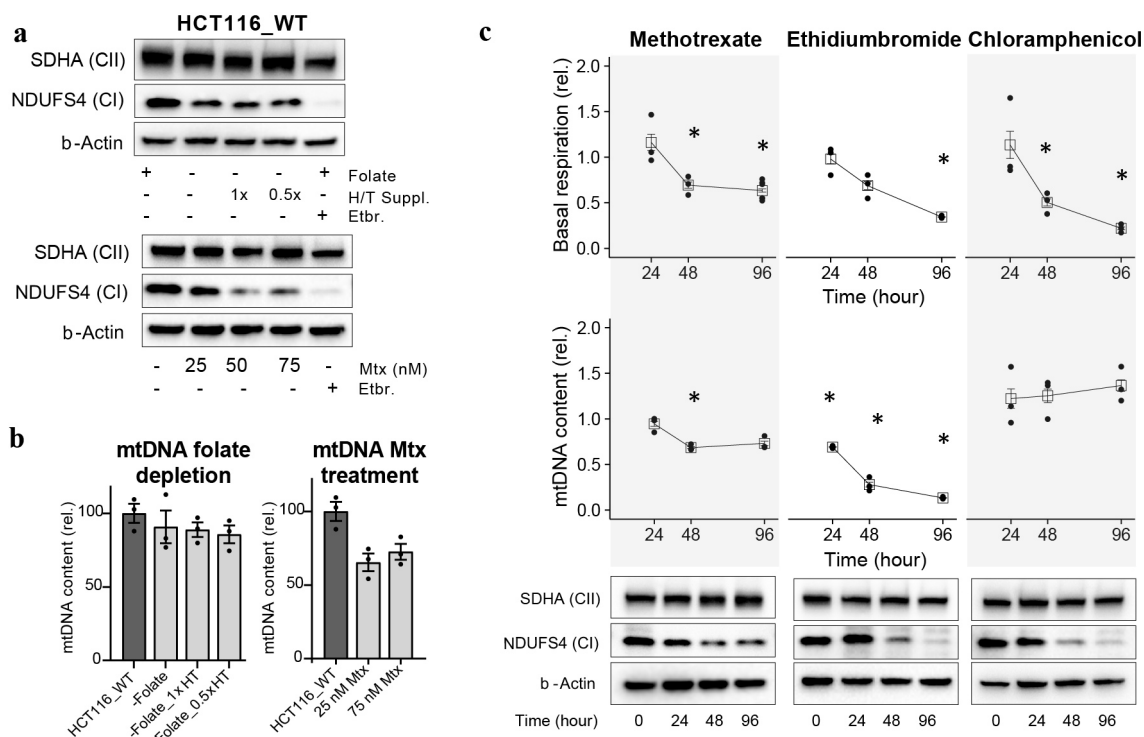
a, Identification of mitochondrial RNA secondary structure based on analysis of the mitochondrial transcript data from the dimethyl sulfate sequencing dataset published previously³⁴. *R* values and Gini differences were calculated to detect changes in nucleotide reactivity between the *in vivo* and denatured condition for the complete mitochondrial transcriptome. Coloured points indicate structured regions given in Supplementary Table 4. **b**, Determination of ribosome stalling sites in SHMT2-knockout HCT116 cell lines. Data points represent individual codons of all 13 mitochondrial protein-coding transcripts. For each codon,

the *y* axis indicates the ribosome counts normalized to the gene median in RPM. The *x* axis indicates the ratio of normalized counts in SHMT2-knockout to normalized counts in wild-type HCT116. Two and three s.d. above the mean of all codons in the genome are indicated by the grey and black dotted line, respectively. Highlighted in red are codons with greater than 2 s.d. **c**, Mapping of AAG and UUG codons from SHMT2 knockout-specific ribosome stalling sites ($>3\ s.d.$) on protein structures. For **b** and **c**, analysis is based on ribosome profiling data in Fig. 3, with two technical replicates of two independent samples. A list of identified codons and mapped AAG and UUG sites is provided in Supplementary Table 5.



Extended Data Figure 9 | Mitochondrial transcript codon occupancy from ribosome profiling of individual patient lines. a, Codon-specific mitochondrial ribosome occupancy ratio (patient/control fibroblasts) in individual patient derived cell lines ($n = 1$ for each individual patient, normalized to mean of $n = 2$ control fibroblast lines). Patients either had nuclear *MTO1* missense mutations (patient A c.[1261-5T>G];[1430G>A], patient B c.[1222T>A];[1222T>A]) or were diagnosed with MELAS and

carry the recurrent point mutation m.3243A>G in the mitochondrial gene for tRNA Leu1 (*MT-TL1*). **b,** Next-generation sequencing of mtDNA mutation load m.3243A>G (*MT-TL1*) in control fibroblasts and MELAS patient cell lines. Each bar shows one biological replicate for control and patient cell lines. Integrative genomics viewer sequencing raw data are shown on the right.



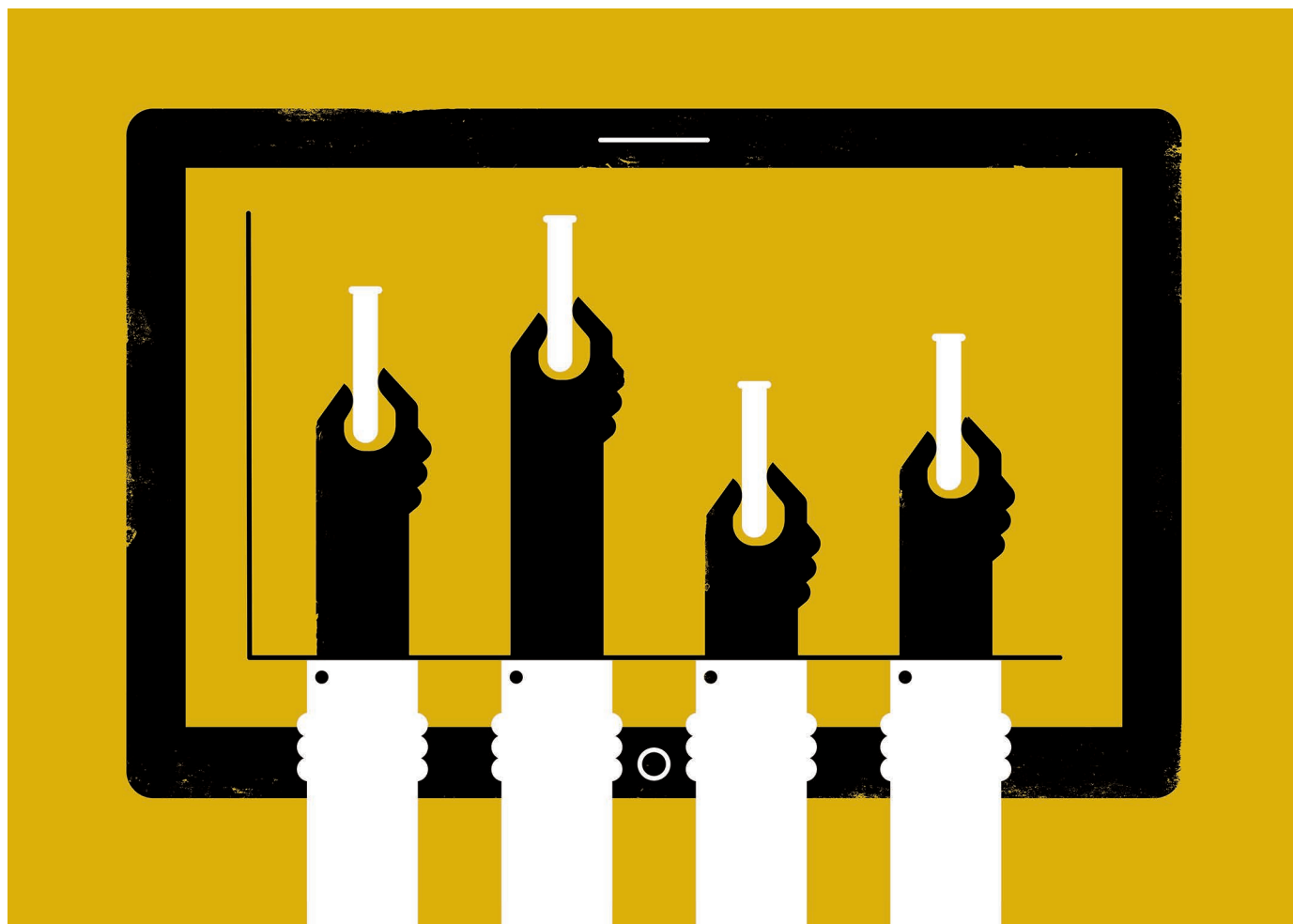
Extended Data Figure 10 | Effects of targeting 1C metabolism on mitochondrial function. **a**, Mitochondrial complex I and II levels after growth in the absence of folate for five passages or in the presence of the indicated methotrexate concentration for 96 h. Ethidium bromide (250 nM) was used as a positive control. **b**, Cellular mtDNA levels in HCT116 cells after folate depletion (with or without 100 μ M hypoxanthine and 16 μ M thymidine (HT) as rescue agents) or in the presence of methotrexate for 96 h ($n = 3$). **c**, To determine whether the decrease in respiration due to methotrexate arises from methotrexate depleting mitochondrial DNA, impairing mitochondrial translation, or a combination, in HCT116 cells we compared the effects of methotrexate (50 nM) to ethidium bromide (250 nM = 100 μ g ml^{-1}), which is

classically used to deplete mitochondrial DNA, and to chloramphenicol (310 μ M = 100 μ g ml^{-1}), which blocks mitochondrial translation. After 48 h of treatment, methotrexate and ethidium bromide both decreased oxygen consumption and DNA content. Importantly, despite ethidium bromide depleting mitochondrial DNA much more strongly, methotrexate had an equivalent effect on oxygen consumption, consistent with the effect of methotrexate on oxygen consumption being in part via mitochondrial translation inhibition. Data are normalized and compared to untreated control (all $n = 3$; except oxygen consumption methotrexate 96 h $n = 6$ and control $n = 4$). Data are mean \pm s.e.m. n indicates the number of biological replicates. $*P < 0.01$, two-tailed Student's t -test (see Supplementary Table 7 for exact P values).

THE FUTURE OF SCIENTIFIC FIGURES

New tools for building interactive figures and software make scientific data more accessible, and reproducible.

ILLUSTRATION BY THE PROJECT TWINS



BY JEFFREY M. PERKEL

As Benjamin Delory started his paper documenting a new way to quantify plant morphology, he realized that one of the figures could pose a problem.

The paper proposes a ‘persistence barcode’ to describe the branching structure of plant root systems¹. The challenge was how to illustrate it.

The barcode’s underlying algorithm “is continuous and dynamic”, says Delory, a

postdoctoral researcher at Leuphana University of Lüneburg in Germany. “And the best solution to show something dynamic is to animate it.”

Scientific figures are typically rendered as static images. But these are divorced from the underlying data, which prevents readers from exploring them in more detail by, for instance, zooming in on features of interest. For genomics needing to cram millions of data points into dense visuals a few centimetres big, this can be particularly problematic.

The same is true for researchers working with computational algorithms. Scientists often post software on open-source repositories such as GitHub, but getting the code to run properly is easier said than done. Reviewers and other interested parties often require extra software and configuration to make the algorithms work.

Some journals now bridge that gap by supporting interactive figures and code. One of those is *F1000Research*, which last year partnered with the computing firm Plotly in ►

► Montreal, Canada, and the Code Ocean platform in New York City. These capabilities, as well as *F1000Research*'s open-access ethos, led Delory and his collaborators to submit their paper there. It was published in January¹.

THE INTERACTIVE PUBLICATION

Interactive graphics that allow readers to delve into a story's underlying data are frequent features on websites such as those of the *New York Times* and *fivethirtyeight.com*, but are less common in scientific publishing.

F1000Research's 'living figures' — interactive charts introduced in 2014 that could be continually updated with new data — were laborious to produce and unscaleable, says senior publishing editor Thomas Ingraham. Plotly lets users build and share visualizations ranging from scatter plots and line graphs to contour plots and maps. The resulting images allow users to zoom in on data, pan across images and mouse-over points to see the plotted values. Student subscriptions start at US\$59 per year. Open-source libraries allow researchers to create free Plotly graphics from R, MATLAB, Python and Julia code.

Code Ocean is free for academics for 10 hours of computation time per month and 50 gigabytes of storage; paid tiers start at \$19 per month. It brings together code, data, results and the computing environment used to execute them in a self-contained 'compute capsule' that replicates the author's computational configuration. Other users can download, modify and run that code either from *codeocean.com*, or through a widget in the paper.

F1000Research has now published six papers with live Plotly graphs and five with a Code Ocean widget. And this year, it plans to add support for interactive protein-protein interaction maps, which are produced using the network-mapping tool Cytoscape.

Researchers need not be put off by the perceived complexity. According to computational biologist Xijin Ge at South Dakota State University in Brookings, who has included interactive Plotly graphs in one of his papers², creating those figures requires just one extra line of code per figure. Tom DeCarlo, a coral researcher at the Oceans Institute and School of Earth Sciences at the University of Western Australia in Crawley, has created six Code Ocean projects for journals including *Paleoceanography and Paleoclimatology* and *Biogeosciences*. "I thought it was really important for scientific communication and reproducibility," he says.

OPEN-SOURCE SOLUTIONS

For those seeking open-source computational alternatives, a tool known as Binder can convert any public GitHub repository containing a Jupyter notebook (documents that interleave text, code and data) or R code into a package that users can run from their browser. Users simply type the notebook repository address into the search bar at *mybinder.org*, and the

program creates a shareable interactive workspace. "It really lends itself to reproducibility and ease of use," says Carol Willing, a Binder project team member at California Polytechnic State University (Cal Poly) in San Luis Obispo.

Such tools also simplify peer review, says Tim Head, a member of the Binder project team in Zürich, Switzerland. Head was frustrated that he couldn't make the software work when asked to review a journal article. "Had they sent me a Binder link, we'd be done by now," he says.

Open-source options also exist for creating interactive images, including Bokeh, *htmlwidgets*, *pygal* and *ipywidgets*. Most are used programmatically, generally within either R or Python code, which is commonly used in science. Coders can, for example, use *ipywidgets* to drop interactive 3D plots, maps and molecular visualizations into Jupyter notebooks. Another option, which is written in JavaScript, is *Vega-Lite*. Because that language is less popular in science, Brian Granger at Cal Poly and Jake VanderPlas at the University of Washington in Seattle developed a Python interface called *Altair* to make it more accessible.

Whereas most of these tools tend to provide functions for specific graph types, *Vega-Lite* and *Altair* are flexible 'grammars' that describe, for instance, how variables map to different visual features, such as colour or shape. They also allow graphs to be linked, such that when users select a region of one plot, the displays of its neighbours update accordingly. "It lets us actually explore relationships in a multidimensional way," says Jeffrey Heer, a computer scientist at the University of Washington whose lab developed *Vega-Lite*.

Two other products let researchers create interactive apps that make use of widgets such as drop-down menus and slider controls to blend data, graphics and code: *Shiny*, made by RStudio in Boston, Massachusetts, for R, and *Plotly's Dash* for Python. They work by transmitting the user's widget actions to a remote server, which runs the underlying code and updates the page.

The resulting apps can make data and tools accessible to researchers who are uncomfortable with programming. For instance, graduate student Tal Galili worked with colleagues at Tel Aviv University to develop a Plotly-based toolbox to build interactive heat maps from uploaded data sets, as well as a *Shiny* interface that runs the code behind the scenes. Mine Çetinkaya-Rundel, a statistician at Duke University in Durham, North Carolina, has built *Shiny* resources for her undergraduate statistics courses to help her to illustrate difficult concepts during lectures.

"It's nice to just pull that up and say, 'okay, now that we've introduced this thing, what happens when we move around the widgets?'" she says.

Publishing such integrations on journal web

pages involves making changes to authoring tools, editorial workflows and infrastructure. It might also involve entrusting scientific data to third parties, who cannot always guarantee their permanence.

To help address this, open-access publisher *eLife*'s Reproducible Document Stack project aims to create an end-to-end tool set for authoring, submitting and publishing documents that are computationally reproducible, says Giuliano Maciocci, who leads product development at *eLife*. The plan is to encapsulate many of a paper's core scientific 'artefacts' — its text, figures, code, data and computational environment — in a single downloadable object, he says. To encourage adoption, the journal is making the stack open source.

MAKING HEADWAY

Several other journals and publishers now support Code Ocean integration, including *GigaScience*, IEEE, SPIE, Cambridge University Press and Taylor & Francis. The *Journal of Cell Biology*'s JCB DataViewer, based on open-source OMERO software, lets readers explore raw microscopy images rather than the processed, compressed files they typically see. A related tool, the Image Data Resource, offers similar functionality for papers published in any journal. *Nature*, too, has published interactive figures, for instance in a paper describing the Encyclopedia of DNA Elements project³. A spokesperson says that the journal is investigating several other options for interactive code and figures. In the meantime, researchers often link to external visualizations from their articles.

As more journals embrace interactivity, the online presentation of scientific information could fundamentally change, representing a win for reproducibility, says Erez Lieberman Aiden of the Baylor College of Medicine in Houston, Texas, who published interactive chromatin interaction maps in a recent *Cell* paper⁴. Static figures are just one perspective on the data. "Informed readers need the ability to draw their own conclusions," he says. "The act of reading a paper in 1974 and the act of reading a paper in 2017 shouldn't be the same act." ■

Jeffrey M. Perkel is *Nature's* technology editor.

1. Delory, B. M. et al. *F1000Research* **7**, 22 (2018).
2. Jung, D. & Ge, X. *F1000Research* **6**, 1969 (2017).
3. The ENCODE Project Consortium. *Nature* **489**, 57–74 (2012).
4. Rao, S. S. P. et al. *Cell* **171**, 305–320 (2017).

CORRECTION

The Toolbox 'A bioinformatics workshop in a box' (*Nature* **552**, 137–138; 2017) erroneously affiliated Robert Gentleman with the Cambridge campus of Harvard University. He was, in fact, at the Harvard T. H. Chan School of Public Health in Boston, Massachusetts.

CAREERS

SPEAK OUT Join with others to make a difference on climate change **p.137**

BLOG Personal stories and careers counsel
<http://blogs.nature.com/naturejobs>

NATUREJOBS For the latest career listings and advice www.naturejobs.com

ILLUSTRATIONS ADAPTED FROM GETTY



COMMUNICATION

Life in the old blog yet

Blogging is still a top way to communicate science to a broad readership, researchers say.

BY ERYN BROWN AND CHRIS WOOLSTON

Allison McDonald has had a lot on her mind lately. She has ruminated on the common mistakes that students make when defending their theses, the identification of the flies that have invaded her office and the plot points of the TV show *Star Trek: Discovery*. But instead of keeping these thoughts to herself, McDonald, a cell biologist at Wilfrid Laurier University in Waterloo, Canada, has posted them on her DoctorAl blog (aemcdonald.wordpress.com).

Science blogs have been around since the early 2000s, and in recent years the 'microblogging' platform Twitter and other social-media channels, which require less time to maintain than does a full blog, threatened to make them obsolete. But some scientists

are keeping the practice alive, and it continues to play a major part in sparking collaborations, conveying crucial information and strengthening scientific communities.

"Blogging isn't for everyone, but it's important that people realize it is part of the many ways scientists talk to each other," says Stephen Heard, an evolutionary ecologist at the University of New Brunswick in Canada and author of the blog *Scientist Sees Squirrel* (go.nature.com/2gk4gf2; tagline: 'Seldom original. Often wrong. Occasionally interesting.')

Studies on the reach and impact of science blogging have refocused attention to the endeavour. In unpublished work, researchers at the Karlsruhe Institute of Technology in Germany surveyed the social-media and scientific-outreach activities of 865 scientists who were

born in 1981 or later. The participants included mathematicians, chemists, physiologists and physicists. Overall, 15% had started a blog, but few updated it with any regularity. "I already knew science blogging wasn't very popular in Germany," says lead author Carsten Könneker, a science-communication researcher who has trained hundreds of young scientists in public outreach. "Blogging is only one digital format for science communication. Scientists who don't make use of any of these formats are missing out on immense opportunities."

The survey uncovered some telling attitudes towards blogs and other forms of science outreach. Nearly two-thirds of respondents said that a lack of time was a 'great obstacle' to any sort of science communication.

But almost 70% agreed that communicating science can help to advance a researcher's

► career, and nearly 90% said that it could help to recruit more bright minds to science.

McDonald had young researchers in mind when she started her blog in 2013. Writing maybe three times a week, she aims to pass on information that could help them to navigate tricky professional waters. “My posts aren’t all epically insightful,” she says. “But the ultimate goal is to take the mystery out of the equation, to level the playing field for people who aren’t aware that there is even a game at play.”

Like McDonald, Heard hopes to inform and encourage younger scientists through his blog. But he also sees benefits to his own career. “I don’t have any evidence that blogging makes it any easier to get grants or to get papers published,” he says. “I have just as many failures now as before. But I have a network of people that I know because they read and comment on my blog posts. There’s a research project on my screen right now that began as a blog post.”

Heard estimates that he averages three to four hours a week working on his blog, but acknowledges that some posts take longer than others. “I’ve spent eight hours writing a single post,” he says. Still, he finds a way to fit blogging into his schedule. “I try to blog at low-productivity times, like when I’m in an airport lounge or waiting for a meeting to start.”

For Heard and others, the investment is worth it. In an October 2017 paper published in the journal *Royal Society Open Science* (M. E. Saunders *et al.* *R. Soc. Open Sci.* **4**, 170957; 2017), he and seven other blogger-researchers analysed the impact of their own ‘science community’ blogs, sites targeting researchers that focus on the culture and business of doing science. The most-read blog in the sample, *Dynamic Ecology*, has a median viewership of more than 40,000 views a month, whereas *Scientists Sees Squirrel* brings in around 10,000 views. Some of the most important impacts are also impossible to quantify. The paper notes that total strangers have walked up to Heard to thank him for a post that offers advice for introverts trying to cope with a conference.

Any study into the reach and impact of blogging will leave some unanswered questions, says Paige Brown Jarreau, a science-communication specialist at Louisiana State University in Baton Rouge who blogs at *From the Lab Bench*. “Blogs are often difficult to define; the ecosystem of online science social-media content is expanding, and platforms are blending into one another,” she says.

Still, blogs clearly have some reach. In a 2017 study that Jarreau co-wrote for *Journalism and Mass Communication Quarterly*, 40 out of 43 randomly selected science bloggers reported getting more than 1,000 views within a few days for a typical post (P. B. Jarreau and L. Porter *Journal. Mass Commun. Q.* <http://doi.org/cjvj>; 2017). For the most part, those clicks were coming from colleagues or colleagues-in-the-making. More than 40% of blog readers

BLOGGING

How to get started

Launching a blog can be daunting, says Stephen Heard, an evolutionary ecologist at the University of New Brunswick in Canada who runs his own blog, *Scientist Sees Squirrel*. Here are some tips:

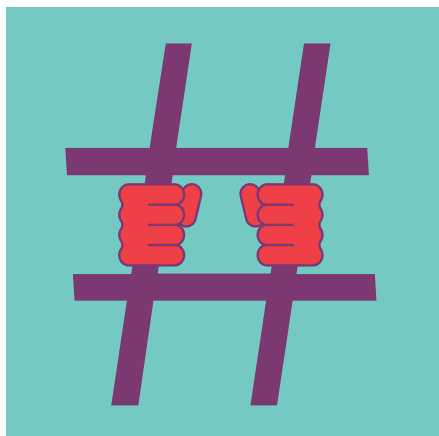
- Choose the right web-hosting service. Heard uses WordPress.com, but notes that there are many other easy-to-use options. For some pros and cons of different platforms, see go.nature.com/2bvaccf.
- Interact with other blogs before launching your own. Simply by commenting on other posts, would-be bloggers can make themselves known to the blogging

community. Guest-posting on an established blog can be another way to get exposure, Heard says.

- Find ways to increase readership. Experiment with different headlines; use strong keywords that are practical and entertaining; and tweet your blog posts or put them on Facebook.
- Don’t get discouraged if page views seem low at first, Heard says, and don’t expect a flood of comments. “A lot of commenting now takes place on Twitter rather than the blog itself,” he says. “Building an audience takes time. But they do build.” **C.W.**

surveyed said that they were already pursuing careers in science, and another 16% said that they were interested in science careers.

For Marc Robinson-Rechavi, a researcher in bioinformatics at the University of Lausanne, Switzerland, one motive for writing his blog was a desire to create a community of bloggers — and readers — in French (most science writing online, including blogs, is in English). Today, Robinson-Rechavi blogs on the French-language platform *Café Des Sciences*, which hosts several dozen bloggers in Europe, North America and Africa.



Terry McGlynn, an ecologist at California State University Dominguez Hills in Carson, credits his popular blog *Small Pond Science* with transforming his career by increasing his visibility. His institution — a teaching-focused school that maintains a relatively low profile among researchers — recently wanted to hire an ecologist. Some of those who were interviewed for the job said that it was McGlynn’s blog that had alerted them to the opportunity. “That was a light-bulb moment for people in my department,” he says. His standing in the blogosphere has helped him in his own negotiations with the university, giving him

the edge he needed to demand more support from his department.

Seeing the positive impact of blogging in his own career, McGlynn has decided to open up the opportunity to others. This spring, he plans to launch *Rapid Ecology*, a site that will feature blog posts from any scientist or science student anywhere in the world who wants to contribute. “There are only three criteria for posts,” he says. “They have to be relevant, they have to have at least some substance and you can’t be a jerk.” He says that 30 scientists have already volunteered to help run the site and contribute occasional posts.

Contributors to *Rapid Ecology* will be limited to one post a month, meaning that anyone could give blogging a try without a huge commitment. “I want students to be able to write posts that will have as much visibility as something that I write in *Small Pond Science*,” McGlynn says.

HASH OUT A FRESH APPROACH

Changes in the online landscape — particularly the social-media boom — have diluted the impact of blogging, argues Jeremy Caplan, director of education for the Tow-Knight Center for Entrepreneurial Journalism at the City University of New York Graduate School of Journalism. With a Twitter or Facebook feed to do the work for them, he says, “people don’t want to keep track of 10, 20 or 30 individual scientists’ blogs”.

And posting to sites such as Medium, Quora and Reddit — ‘hangout’ sites where researchers or any subgroup can post ongoing ‘threads’, or conversations on a single topic — is a way to publish without the burden of maintaining a blog (see ‘Blogging’). Responding to the constant need for new content, say bloggers, can take tremendous discipline. “Writers lose steam and decide to park their store in someone else’s mall,” Caplan says.

As Jarreau sees it, social-media platforms don’t supplant blogging, they feed it — giving

writers a place to develop and test ideas that they might later incorporate into a lengthier post, and directing readers to the detailed content they want. “Discovery of science blogs is increasingly through social media,” she says.

RISKS AND BENEFITS

Blogging does have potential pitfalls. For a start, it is not likely to make anyone wealthy. “It’s probably not worth doing it for the money unless your audience is huge,” says McGlynn. Small Pond Science, which has had more than 570,000 visits in total, doesn’t take ads. But even if it did, McGlynn has calculated, he’d probably clear only US\$10,000 to \$20,000 a year.

Academic colleagues might think that blogging is a waste of time or damaging to a career. “Some people say blogging and social media are distractions and will hurt you on the job market because it demonstrates that you’re not serious,” McGlynn says. When Robinson-Rechavi started blogging in 2010, he signed his posts using only his initials, unsure how people would react — even though he already had tenure and faced little risk. He thinks that his colleagues don’t understand why he blogs and are indifferent to his posts. Yet administrators at his university consider his blogging a useful forum for communicating ideas. “I think they like that I’m doing it,” he adds.

It’s worth considering the inherent risk in putting one’s name, face and ideas on the Internet. McDonald says that bloggers — and particularly women — need to think carefully before they post, because online visibility can expose writers to abuse.

Still, McDonald keeps at it, happy to be involved in broader conversations about teaching, biology, women in academia — and *Star Trek*. When she came up for tenure, she discussed her blog in her application. “This is part of my outreach and advocacy work for diversity in science,” she says. Blogging helps her to take her research into the world, a goal that she believes is crucial for scientists.

“We hear all the time about the decline of blogging,” Heard says. But he has no intention of quitting and will continue to spread the word about its benefits. “I hope that those who are on the fence — those who think it might be for them — can be encouraged to give it a go.” ■

Eryn Brown is a writer and editor in Los Angeles, California. **Chris Woolston** is a freelance writer in Billings, Montana.

COLUMN

Make yourself heard

Researchers who want to ‘do something about it’ can join with others to effect change, says **Sarah Hamylton**.

The impacts of climate change are real. Plant and animal habitats are changing, glaciers are melting and heatwaves and floods are becoming more frequent. All this causes me to question the utility of my work as an environmental scientist.

Reports that two mass coral-bleaching episodes in 2016 and 2017 had killed around half the coral on the Great Barrier Reef stopped me short. Having spent the past decade modelling the impacts of climate change on coral reefs, I feel as if much of that work is now futile.

Environmental scientists are calling attention to changes in the natural world that are driven by carbon emissions from burning fossil fuels. In doing so, we speak of the ‘grief’ of climate science, using words such as ‘demoralizing’, ‘conflicted’ and ‘deep sense of worry’. Charlie Veron, a world authority on coral and former chief scientist of the Australian Institute of Marine Science, told the Australian Broadcasting Corporation’s Radio National in 2016: “I am someone who can actually do something about it. I am someone who is listened to and I have made a difference. And so I have to keep on doing that. It’s not as if I can say, ‘to hell with it’, and go and do some gardening.”

We have a responsibility to lead change. This responsibility raises questions, such as: how do scientists cope with the emotional burden of their knowledge? And how can these emotions galvanize us into action?

In 2017, I joined Homeward Bound, a global environmental-leadership programme for women in science that launched in 2016. Each year, the programme coaches up to 150 women for 12 months, culminating in a 3-week voyage to Antarctica, where female scientists develop their confidence and strategic vision for acting together on climate change. The programme has focused my attention on projecting my voice as an environmental scientist.

In 2015, I became a councillor of the Australian Coral Reef Society (ACRS), the world’s oldest society for protecting Australia’s coral reefs, which has a track record of calling for change. That year, we wrote submissions and reports on behalf of more than 300 concerned scientists in what became known as the ‘coal versus coral’ war. The Great Barrier Reef Marine Park Authority had approved a proposal to dump 3 million cubic metres of dredged sediment from Abbot Point, a huge coal port in northern Queensland, into the



Great Barrier Reef World Heritage Area.

This would have been an environmental disaster, with plumes of sediment compromising marine life. The authority reversed its decision when the ACRS made its views known alongside those of conservationists, tourism operators, grassroots organizations such as GetUp! and the indigenous climate group Seed. It was immensely satisfying to be part of this endeavour. To keep up the pressure, we sent a letter last August on behalf of the ACRS to Australia’s prime minister, Malcolm Turnbull, urging immediate action to curb carbon emissions.

I have also begun to explore how interdisciplinary approaches weave together different practices to create powerful ways of communicating the science of climate change. Last September, I became an unlikely ‘artist in residence’ at the Bundanon Trust in Illaroo, Australia, which supports creative work that emphasizes the value of landscapes. I am working with artists and a social scientist to untangle how interdisciplinary approaches saved the Great Barrier Reef from mining in the 1960s — and whether such approaches can help scientists to save it again.

Emotional conflicts around climate change have prompted me to revisit the reasons I became an environmental scientist. I am now using forms of expression that resonate with my personal values and add scientific authority to the argument for resisting the coal industry. How will you lead the change you want to see? ■

Sarah Hamylton is a senior lecturer in geographic information sciences at the University of Wollongong, Australia.

SEVEN POINT TWO

Contact has been made.

BY MARISSA LINGEN

The astronomers gathered behind the podium, beaming proudly. They had never expected to have so much of the mainstream press at one of their press conferences; most of them had never expected a press conference in the first place. The bright lights of the video cameras made them squint, their smiles slipping into grim desperation even in their moment of triumph.

Cathy Biyu Li had been chosen as the spokesperson, and she spoke. “Good evening. We would like to confirm that we have decoded the broadcast from the unexpected object in the Orion Nebula. Our early reports that it was highly structured have been verified by extensive testing. We now feel extremely confident that this was a communication by an intelligence that was not terrestrial in origin. The communication was as follows —”

“Aliens?” shouted the BBC reporter. “Are you really telling us you have a code from aliens?”

“Yes, that’s exactly what I’m saying,” said Dr Li. “If you’ll let me finish, I can tell you what they were saying.”

The reporters took a moment to stop shouting questions and comments. They settled to a buzz, then to silence, as Dr Li waited.

“First we had some number sequences, setting the communication with primes and triangular numbers and perfect numbers. This is particularly encouraging because it shows that we have similar concepts of mathematics.

“We have several key ratios that appear in nature: an approximation of twice π , which shows up in calculating a circumference. That one went out to 20 digits and could not be mistaken for a random number or for something else. We found the ratio of the mass of the proton to the mass of the electron, that one out to ten digits, again not something that would show up randomly. That number, you will note, requires a fairly high degree of scientific measurement, and yet is not dependent upon the units used for that measurement.”

The science writers nodded and scribbled quick notes, in their element. They had planned for this, and it was more or less

what they expected. The mainstream press squirmed and tried to pay attention, but they had been hoping for more portraits of aliens, and fewer strings of numbers — especially with lots of complicated science facts attached to them. The list of numbers droned on.

“... and seven point two,” said Dr Li. “Thank you all for coming, we’re so glad you could —”



“What’s seven point two?” asked the reporter from the *Times*, who was not afraid of looking stupid. “You explained all the rest of them.”

“We’re ... not really sure yet. All the others were meaningful. We’re looking into it.”

Pandemonium among the reporters.

Six months later, ‘looking into it’ did not begin to cover the worldwide efforts to ascertain the importance of the constant seven point two. Funding for fundamental research skyrocketed. Humanity had been handed a key to the Universe, if only they could find the lock it fit within.

At first the big winners were in physics and chemistry, but over time the planetary scientists/geologists and biologists made their claims known. Surely, they argued, there would be some universality of life processes, some press of rock and sand that remained itself regardless of particulars of atmosphere and temperature. Unitless constants were rare and

powerful things in the world, but every university’s science departments were

willing to be richly endowed in order to find more of them.

Or at least to pursue them indefinitely.

Meanwhile, Cathy Biyu Li and the other astronomers had kept listening to the signal from the unexpected object in the Orion Nebula. It did not seem to be redshifted or blueshifted compared to the objects around it. It continued to emit a signal, but the signal repeated. Over and over again, so there

could be no mistake: here are our constants, here are the things we know about the Universe, the tiniest pieces, the repeating parts.

And seven point two.

One of the other astronomers, Dr Jorge Estrada, found Dr Li signalling back to the Orion Nebula one night, late, when no one was around. The same patterns, the same numbers. Maths, physics, the Universe as we know it. And then.

“You’re not telling them seven point two. What if they — what if they think we’re not sentient, what if they — what was that?” he demanded.

Cathy Biyu Li grinned. “Twelve point nine.”

“What are you doing?”

“Aliens are testing us for all sorts of things, Jorge. Curiosity, mathematical ability, advancement in the physical sciences.”

“Yes, of course, we’d be looking for all of those things.”

“And maybe seven point two is there to get us looking farther. To stimulate us, even if there’s nothing to it.”

Jorge nodded; this was a popular theory among the astronomers.

“But maybe,” Cathy continued. “Maybe they’re testing for something else. Maybe it’s their little joke. Children love randomness. If you want to make a baby laugh, do something funny. Unexpected. Seven point two? Hah, that’s a good one. You want to hear another one? Twelve point nine!”

She chortled. He stared at her blankly. “I don’t get it.” He wandered off, shaking his head angrily and muttering about the committee.

“No,” she whispered. “But in another couple of thousand light years, someone else is going to get a good laugh out of it.” ■

Marissa Lingen has published more than 100 short stories in venues such as *Analog*, *Lightspeed* and *Tor.com*.

ILLUSTRATION BY JACEY